Cognitive Theories and Forensic Applications: The Pupillary Correlates of Familiar and Unfamiliar Face Processing

Camilla Elphick

Submitted for the degree of Doctor of Philosophy in Psychology

School of Psychology

University of Sussex

June 2018

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Portions of the thesis have also been published in the Psychology Journals listed below:

Reported in Chapter 4: Elphick, C., Pike, G. & Hole, G. (2011). You can Believe your Eyes: Measuring Implicit Recognition in a Lineup with Pupillometry. Under review at the Journal of Experimental Psychology: Applied (http://www.apa.org/pubs/journals/xap/)

Portions of this thesis were carried out in collaboration with others:

Reported in Chapter 3: Dr Matthew Fysh contributed to this paper by testing participants at the University of Kent. The design, execution and reporting of all experiments was conducted by the author of this thesis and supervised by Dr Graham Hole.

Acknowledgements

First, I want to express my gratitude to my enthusiastic supervisor, Graham Hole. He took me on as an unprepared and overwhelmed student, but has made my Ph.D. a life-changing experience. Without his cheerful and encouraging presence, his insight and integrity, his sensitivity and support, I doubt I would have had the resilience to finish. I thank Graham wholeheartedly, not only for his tremendous academic and emotional support, but also for making the experience so much fun. I will miss the laughter and discussions about the Lego Movie, that have been memorialised in this thesis.

Profound gratitude also goes to Sam Hutton, who has been a kind and supportive advisor. I am particularly indebted to Sam for his constant advice when building eye-tracking experiments and responding to my plaintive cries for help. Above all, I would like to thank him for helping me to build an experiment and lending me his expensive eye tracker (to use in torrential rain on Brighton Pier) on his wedding day!

I am also hugely grateful to Graham Pike, for sharing his stimuli so willingly, for his support in my eyewitness research, and for welcoming me to his lab group. I would also like to thank Matt Fysh for his incredible contribution to the research in this thesis. I could not have done it without his kindness, hard work and positivity.

Special mention goes to Sybella Kirkbride, who was the first of my tutors to encourage me with her excellent teaching, and who was kind enough to provide me a reference to start the Ph.D. Similarly, I am very grateful to Roberto Caldara, who gave me invaluable experience and a reference. Without him taking the time to help, I would not have had the courage to apply for a Ph.D. Thank you Janka Bryjova for helping me to connect with him. I also want to thank Rachael Chivers and Maddie Atkinson for taking an interest in my research and helping me with participant recruitment.

I would also like to thank and apologise to my wonderful friends and relations for putting up with me all this time, whilst being ignored. I am truly grateful to your support and patience. I am looking forward to spending more time with you.

To my parents, I owe so much. They have encouraged and supported me throughout, with unfailing confidence in my ability. Thank you for making the experience so much easier by always being there to help and to celebrate the small milestones.

Finally, I cannot express how lucky I feel to have had my wonderful husband by my side. He encouraged me to take my first psychology module ten years ago, and has willingly made numerous sacrifices so that I could pursue this dream. He has endured my highs and lows without complaint, and has never doubted me. I dedicate this thesis to him (and my dogs, who have warmed my lap and feet throughout).

Four Feet

Day after day, the whole day through
Wherever my road inclined,
Four feet said, “I’m coming with you”
And trotted along behind.

(Rudyard Kipling)

UNIVERSITY OF SUSSEX

Camilla Elphick

Doctor of Philosophy in Psychology

Cognitive Theories and Forensic Applications: The Pupillary Correlates of Familiar and Unfamiliar Face Processing

Summary

This thesis used pupillometry to investigate whether pupils respond differently to faces that differ in familiarity. We aimed to see whether pupillometry measures cognitive processes involved in face processing, and whether it could be applied forensically.

We started by evaluating three explanations for pupillary changes that occur when processing faces. The first was cognitive load (mental effort), because faces that have only been seen briefly are more difficult to recognise than well-known faces. The second was cognitive engagement (interest), because faces contain socially-important information. The third was memory strength (forensically applicable), as eyewitnesses have to recall a perpetrator's face in an attempt to identify them if they appear in a lineup. While pupillary responses reflected cognitive engagement to some extent, cognitive load best accounted for decreasing pupil sizes when learning new faces, and memory strength explained the pupillary

changes seen in lineups. The theories all had some influence on pupil sizes, but their influence varied according to context, saliency, and the task at hand.

Then we investigated whether pupillometry measured implicit recognition of a perpetrator in a lineup, and found that it did. Pupil sizes reflected memory strength in participants who believed their memory to be strong: there were differences in pupil sizes (between looking at the perpetrator and the distractors) in participants who identified him, but not in those who did not. The pupillary responses of participants who 'guessed' indicated that they were indeed guessing. There were no pupillary changes when the perpetrator was not in the lineup, even when participants misidentified a distractor. We concluded that pupillary responses are independent of explicit identification responses, and could be used forensically to support traditional measures of eyewitness identification and credibility.

Contents

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA – Analysis of Variance
CCTV – Closed Circuit Television
CLT – Cognitive Load Theory
DV – Dependent Variable
ERP – Event-related Potential
FRU – Face Recognition Unit
GFMT – Glasgow Face Matching Task
HSF – High Spatial Frequency
IAC – Interactive Activation and Competition model
ID – Identification
ISI – Interstimulus Interval
LSF – Low Spatial Frequency
NSF – Normal Spatial Frequency
OAB – Own Age Bias
OGB – Own Gender Bias
ORE – Other Race Effect
PIN – Person Identity Nodes
RK(G) – Remember Know (Guess)
RT – Reaction Time
SD - Standard Deviation
SE - Standard Error
SEM - Standard Error of the Mean
UK – United Kingdom
US – United States
VIPER – Video Identification Parade Electronic Recording

CHAPTER 1. COGNITIVE THEORIES AND FORENSIC APPLICATIONS: THE PUPILLARY EFFECTS OF FAMILIAR AND UNFAMILIAR FACE PROCESSING – OVERVIEW

1.1. Why Is It Important to Study Differences Between Familiar and Unfamiliar Face Recognition Processing?

Faces reveal considerable amounts of information about individuals, such as race (e.g. Chiroro & Valentine, 1995), age (e.g. George & Hole, 1995), gender (e.g. Wright & Sladden, 2003), attractiveness (e.g. Chatterjee, Thomas, Smith, & Aguirre, 2009), health (e.g. Fink, Neave, Manning, & Grammer, 2006), emotion (e.g. Calder, Young, Keane, & Dean, 2000), and identity. Humans are social creatures, so being able to determine these characteristics and states is important. Indeed, infants seem to be programmed to take an interest in faces over other objects (Mondloch et al., 1999), and are able to distinguish between faces soon after birth, preferring their mother's face over other faces (Bushnell, 1998). People are generally able to make relatively accurate classifications of faces with just a glance, even when they are unfamiliar. However, while *recognising* the faces of people who are highly familiar is also easy for most people, recognising those of people who have only been seen briefly before is more difficult, suggesting that there are some differences in the processing of familiar and unfamiliar faces for identification purposes (see Johnston & Edmonds, 2009, and Jenkins & Burton, 2011, for reviews).

The importance of studying the differences between familiar and unfamiliar face processing is demonstrated in two main ways. First, people with prosopagnosia, who have severe deficits in recognising familiar faces, can suffer psychosocial and lifestyle consequences such as avoidance of social interactions, anxiety, loss of confidence, and loss of employment opportunities (e.g. Yardley, McDermott, Pisarski, Duchaine, & Nakayama, 2008; Dalrymple et al., 2014). Second, eyewitnesses are expected to recognise unfamiliar faces at a level associated with familiar face recognition, but misidentifications can result in life-changing wrongful convictions. It is reported that over 70% of wrongful convictions that have been overturned in the US were related to eyewitness misidentifications (The Innocence Project, n.d.).

1.1.1. An Introduction to the Differences between Familiar and Unfamiliar Face Processing

Despite considerable research investigating face recognition, certain aspects remain poorly understood. For instance, the appearance of a face changes when seen from different viewpoints or with different lighting conditions. In a series of experiments, Megreya and Burton (2006, 2007, 2008) found that people can accommodate these differences (within-face variability) much better if they know the faces than if the faces are unfamiliar to them. The researchers either showed participants two different images of faces in pairs (and asked them to decide whether the images were of the same person or of two different people), or a target face, shown with ten other faces (and asked them to decide whether a different image of the target face was also present among the ten other faces). They found that participants had no difficulty individuating familiar faces shown from different viewpoints or with different lighting, but performed poorly with unfamiliar faces, indicating that familiar and unfamiliar face recognition use distinct

processes. However, little is known about *how* people account for changes in viewpoint or lighting when recognising an individual face, or how they individuate between faces of different people.

Since the 1970s with the work of researchers like Ellis, Shepherd, and Davies (1979), there have been many investigations into the underlying processes involved in face recognition (Collishaw & Hole, 2000), how unfamiliar faces become familiar (Zimmermann & Eimer, 2013), factors that affect face recognition (Meissner & Brigham, 2001; Deffenbacher, Bornstein, Penrod, & McGorty, 2004) and eyewitness identification issues (see Wells & Olson, 2003 for a review). This thesis will go through each of these, and discuss relevant theories.

1.1.2. The Processes underlying the Differences between Familiar and Unfamiliar Face Processing

The importance of facial features as cues to face recognition has been studied extensively (e.g. Goldstein & Mackenberg, 1966; Smith and Nielsen, 1970), and forms the basis of many recognition procedures in e.g. police work (Frowd et al., 2005). However, the contribution of featural processing to face recognition has long since been overshadowed by theories suggesting that face recognition involves "holistic" and "configural" processing. Some studies indicate that featural cues may be more important than was originally thought (Cabeza & Kato, 2000). However, the distinction between these different processing types has been blurred by conflicting and overlapping definitions and terms, methods and measurements (Richler, Palmeri, & Gauthier, 2012), which have delayed progress in research. This thesis will outline three processing types:

featural, holistic, and configural. The definitions used in these outlines will be used in the rest of the thesis.

1.1.2.1. Featural Processing

Featural processing describes processing individual features to recognise a face. It has also been referred to as "analytical" (e.g. Anaki, Boyd, & Moscovitch, 2007), "part-based" or "piecemeal" processing (e.g. Hole, 1994; Collishaw & Hole, 2000). Although definitions vary, all these terms share the notion of breaking down a face into its constituent parts. Techniques for testing featural processing include either removing features (Young, Hay, McWeeny, Flude, & Ellis, 1985) or changing features (e.g. Tversky and Krantz, 1969), to see whether these features are necessary for face recognition. Part/whole tests (e.g. Tanaka & Farah, 1993) are also used, to see whether the feature is recognised better as part of a whole face, or in isolation. However, the problem with featural processing theories is deciding what constitutes a "part".

1.1.2.2. Configural Processing

Configural processing refers to processes involving the encoding of fine-grain spatial information. Although this is often thought of in terms of "holistic" processing, it need not involve the whole face. This type of processing has also been referred to as second-order processing (e.g. Sandford & Burton, 2014), configurational (Hole, 1994; Collishaw & Hole, 2000), or relational processing, and has even been used to describe a set of processes (e.g. Maurer et al., 2002). Techniques used to test configural processing include using composite face tasks (where the top half of a face is combined with the bottom half of another to make it look like a ‘new’ face) inversion (where faces are presented upside-down) (Maurer, Grand & Mondloch, 2002), slanting (Busey, Brady, &

Cutting, 1990), and distorting, stretching or squashing the face (Hole, George, Eaves, & Rasek, 2002), to see whether recognition is affected when the configural information is disrupted. Other studies present faces in fine grain detail (using high spatial frequencies) or coarse detail (using low spatial frequencies), to see which aspects of the face are more important for recognition (e.g. Cheung, Richler, Palmeri, & Gauthier, 2008). However, one problem with configural processing theories is differentiating between configural and featural processing. For example, eye-separation could be considered a form of configural processing or be referred to as a "feature".

Fig. 1. Example of High spatial frequencies (HSF), low spatial frequencies (LSF) and normal spatial frequencies (NSF) (taken from Feusner et al. 2012).

1.1.2.3. Holistic processing

Holistic processing is a form of processing that uses information from the entire face, not just a localised region. The term has often been used interchangeably with "configural", which can be confusing. It was first proposed by Sir Francis Galton (1883), who described it as perceiving the face as a whole (a Gestalt), rather than as separate features (cited in Goffaux & Rossion, 2006, p.1023). Subsequently, there has been confusion over the definition of holistic processing, as it has been shown that inverting

faces affects both gestalt processing and the ability to judge configurations (Goffaux & Rossion, 2006). The holistic hypothesis thus presents processing as the integration of featural and configural information. Holistic processing can also be tested using composite face tasks, (e.g. Richler, Mack, Gauthier, & Palmeri, 2009), inversion (e.g. Avidan, Tanzer, & Behrmann, 2011), part-whole tasks, (e.g. Tanaka & Farah,1993; Van Belle, De Graef, Verfaillie, Busigny, & Rossion, 2010) and by manipulating spatial frequencies (e.g. Cheung, Richler, Palmeri, & Gauthier, 2008).

However, Collishaw & Hole (2000) demonstrated that people might not rely on just one form of processing, but use the appropriate process for the information available. Also, the cues provided by gender, age, race, attractiveness, distinctiveness, voice, facial expressions, gestures, weight, hairstyle, context etc. are probably used in similarly adaptive ways, depending on which are available. It might be possible to achieve successful face recognition by eliminating faces that do not fit with the cues provided until a small selection of possible faces remain, which require fine-grained analysis, using Bayesian-type hypothesis testing (Fischhoff & Beyth-Marom, 1983; Balas, 2012). In short, it appears that face recognition is a flexible process, where people can use whatever featural, holistic, configural, or cued information that is available to them.

1.1.3. Theoretical Models to Explain the Distinction between Familiar and Unfamiliar Face Processing

Bruce & Young (1986) proposed a framework that accounted for perceptual and cognitive processes involved in face processing, including face recognition. It contains four face recognition units: pictorial, which is a description of a static image (like a photograph); structural, which is a more abstract visual representation of a familiar face

that can mediate recognition when the face is seen from a novel viewpoint; visually derived semantic, which uses visual cues to semantic information to make assumptions about the face (such as age or gender); and identity-specific semantic, which refers to known information about a familiar face (such as occupation). It also contains three sequential identity recognition stages. The first stage describes face recognition units (FRUs) that store visual structural face descriptions. The appropriate FRU becomes activated when a view of a face is recognised, stimulating the person identity node (PIN) in stage two. The PIN accesses identity-based semantic information, which can also be activated by non-face cues such as voice. The third stage is naming the person. However, the model did not explain how familiarity judgments were made (Bruce et al., 1992).

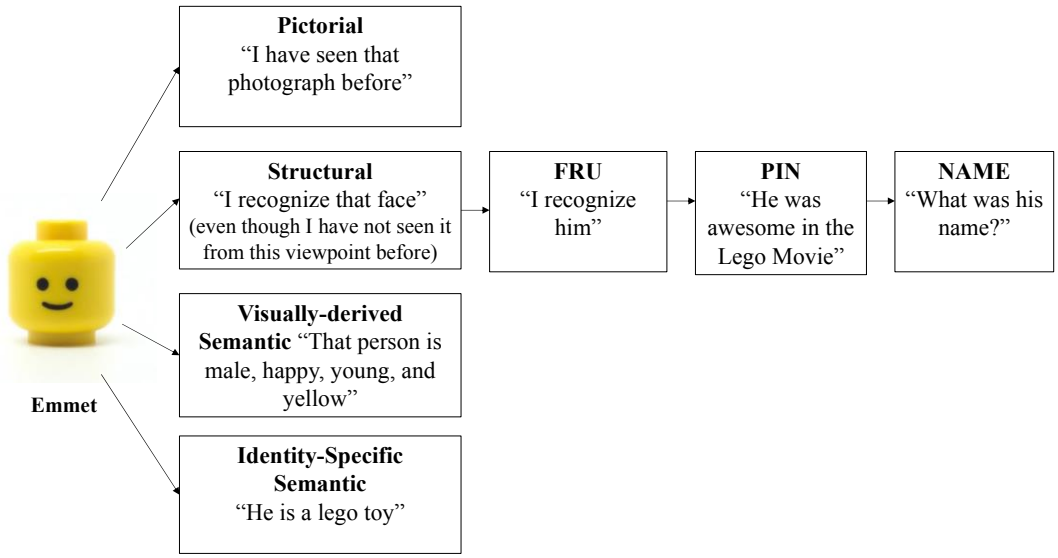


Fig. 2. A simplified representation of Bruce and Young's (1986) framework for processes involved in face recognition

Bruce et al. (1992) expanded this model with a connectionist Interactive Activation and Competition model (IAC) that was originally an attempt at a computer

version of the original model. Later, Burton, Bruce, and Hancock (1999) combined cognitive and perceptual components in a new model. These models are useful for modelling familiar face recognition, but none can explain in detail how faces are learnt.

Hancock, Bruce and Burton (2000) felt that perceptual models focus mainly on unfamiliar faces, which are generally only processed according to visual information, while cognitive models largely focus on familiar faces, which can be processed according to known semantic information (such as occupation). Therefore, they proposed a new model with a front-end image-processing system (that deals with perceptual information), and a connectionist system (that deals with cognitive identification). Hancock et al. (2000) believed that modifications of models such as this could eventually clarify how unfamiliar faces are learnt. However, the task is difficult as the distinction between "familiar" and "unfamiliar" is not clear-cut. This is partly because there are so many ways to experience faces.

1.1.4. What constitutes familiarity?

1.1.4.1. Famous Faces

Many studies use famous faces as stimuli. These are sometimes "iconic" photographs, and probably test pictorial recognition (of the image) rather than face recognition (Carbon, 2008; Burton, 2013). Familiarity with these faces is also generally limited to two-dimensional images (e.g. films, television or magazines), which is not the way we typically become familiar with people in real life. Different images of an actor can also differ from each other more than images of a person we know personally, as actors are employed to portray different characters with different appearances (and accents). This is not typical of personally-familiar face recognition in the real world.

Thus, recognising the faces of famous people is probably different from recognising the faces of people that are known personally.

1.1.4.2. Personally-Familiar Faces

Some studies use personally-familiar faces, as exposure to them is extensive, often spanning years, with multiple views from different angles, in different lighting, with hairstyle and weight changes, and changes associated with health, ageing etc. People also get to know personally-familiar faces as they move in multi-dimensional space. As result, people form stable face representations (mental abstractions) of personally-familiar faces (Hancock, Bruce, & Burton, 2000). These are sensitive to differences between faces of different individuals (between-face differences), but can also accommodate fluctuations in a person's appearance that occur due to different lighting, viewpoints etc. (within-face differences) (Johnston & Edmonds, 2009). These representations are far more robust than those of experimentally-learnt faces. However, using personally-familiar faces can be problematic in an experiment, as faces that are highly familiar to one person may well be unfamiliar to another.

1.1.4.3. Experimentally-familiar Faces

Experience with faces that are familiarised in experimental paradigms cannot produce the robust representations associated with personally-familiar or famous faces (Tong & Nakayama, 1999). However, tests using familiarised faces suggest that relatively little exposure to new faces can be sufficient for them to appear familiar. For example, Clutterbuck and Johnston (2005) found that although familiar faces were matched more quickly than new or newly-learnt faces, newly-learnt faces were matched more quickly than completely new faces. These studies therefore suggest that using familiarised faces

can be useful for understanding how quickly faces become familiar and how exposure can mediate learning. Familiarisation studies also have the advantage of reducing lexical, episodic or semantic memory associated with personally-familiar or famous faces. This makes it easier to draw conclusions about e.g. how much exposure is necessary for the early stages of face learning.

1.1.4.4. Own Faces

Own face images are also sometimes used in experiments. Previous research suggests that they require more effort to process and may be treated as more unfamiliar than other familiar faces (Brédart, 2003) because people generally only see themselves in the mirror (and faces are not entirely symmetrical). However, nowadays, many people see frequent images of their own face on mobile phones or on social media (see Senft & Baym, 2015, for a review), so they should have an intimate knowledge of their own (veridical) face from multiple angles *and* their face seen in the mirror (mirror-reversed). In short, nowadays, people should be more familiar with their own face than they are with any other face. Research also suggests that self-relevant stimuli are important, and this extends to own face images (Kircher et al., 2001; Tacikowski & Nowicka, 2010). This interest in one's own image is demonstrated in the selfie phenomenon, where doctored images are shared as idealised representations of the self (Murray, 2015). Therefore, own face images now provide stimuli that should be both highly familiar in veridical and mirror-reversed format, and particularly engaging to the person viewing them. Thus, the use of own face images can be problematic as the rise in technology has dramatically changed the ways that own face images are accessed. The results are also difficult to generalise as own face images are different for each participant.

1.1.4.5. Unfamiliar Faces

Research has revealed that familiar faces are easy to recognise, while it is much more difficult to recognise a face that has only been seen briefly before. One explanation for this is that representations of familiar faces can be applied flexibly to familiar faces seen in different conditions, while representations of unfamiliar faces rely upon on poor or fragmented information that make them harder to recognise when seen in different conditions (Hancock et al., 2000). This explains why familiar face recognition is good even when image quality is poor, and why unfamiliar face recognition is poor even when image quality is good. Burton, Jenkins, and Schweinberger (2011) suggest that familiar face processing is based on abstract structural codes: when a familiar face is seen, its characteristics can be matched to its stored representations, even when the face is seen under novel conditions. However, unfamiliar face processing is based on pictorial codes that are less flexible and make recognition in different conditions more difficult. Zimmermann & Eimer's (2013) research supports this distinction, as they found that familiar face recognition is possible from multiple views, while unfamiliar face recognition is more view-dependent. These findings all suggest that poor unfamiliar face recognition might be related to limited and inflexible information about the faces.

Research into the role of movement in face processing also supports this view. Knight & Johnston (1997) found that watching videos of moving faces aided recognition. They suggest that this is in part due to providing three-dimensional and characteristic information. Lander & Bruce (2000) found a similar advantage for famous faces, and Lander & Bruce (2003) found that motion improved unfamiliar face learning. They suggest that this is related to increased attention to socially-important facial movement. Xiao, Quinn, Ge, & Lee (2012) concluded that motion affects featural rather than holistic processing, suggesting that featural processing is important to face recognition. Overall, it is likely that movement helps with face recognition, as it provides additional

information about the structure and characteristics of the face seen from multiple viewpoints, allowing for the development of dynamic representations (Pilz, Bülthoff, & Vuong, 2009).

Thus, the distinction between familiar and unfamiliar faces is not simple. Unfamiliar faces can refer to those that are entirely novel or those that have been seen briefly before, yet experimental paradigms often compare faces that have only been seen briefly before to completely novel faces. As for faces that are generally agreed to be familiar, there are also differences: famous faces are generally known two-dimensionally, but personally-familiar faces are experienced contextually and three-dimensionally. Finally, own face recognition is difficult to categorise, as recent advances in technology mean that we have more familiarity with our own faces than ever before. Therefore, the question about how much experience with a face gives rise to a sense of familiarity has not been answered definitively. One theory suggests that faces lie on a continuum of familiarity (Rhodes, 1985): recognition becomes easier as representations become more robust and flexible, until faces can be recognised even from a novel or poor view. This is because increasingly robust (abstract) representations allow people to account for within-face variability that is a consequence of lighting or viewpoint etc., and to separate this from between-face variability that is a consequence of different faces.

1.1.5. Aspects that Affect the Recognition of Unfamiliar Faces

When it comes to recognising unfamiliar faces, it seems that to the eye of the beholder, not all faces are created equally. A large body of work has found that groups of face types that differ from the beholder in terms of appearance, are recognised less successfully than those that resemble the beholder.

For instance, race has been found to moderate face recognition: other-race faces are less easy to recognise and remember than own-race faces, and they take longer to learn than own-race faces, a phenomenon known as the Other Race Effect or ORE (e.g. Cross, Cross, & Daly, 1971; Valentine & Bruce, 1986; O'toole, Deffenbacher, Valentin, & Abdi, 1994; Megreya, White, & Burton, 2011; Meissner, Susa, & Ross, 2013). For example, Michel, Rossion, Han, Chung, & Caldara (2006) found that own-race faces are processed more holistically and with greater accuracy than other-race faces. Knowing more about this phenomenon might clarify how faces are learnt and what might improve learning. For example, DeGutis et al. (2011) found that configural training with *own-race* faces improves *other-race* face processing.

A similar effect is found when looking at other-*age* faces, which appear to be processed less efficiently than own-age faces, a phenomenon known as the Own Age Bias (OAB) (see Rhodes & Anastasi, 2012 for a review). For example, Anastasi & Rhodes (2005) tested older adults and children and found that both groups recognised own-age faces more easily than other-age faces. Konar, Bennett, & Sekuler (2013) also found (in a study that only used young faces as stimuli) that older people processed faces less effectively and more holistically than younger people.

1.1.6. Theoretical Models to Explain the Effects of Race and Age on Unfamiliar Face Processing

Valentine (1991) proposed a framework that conceptualises faces as points in multi-dimensional space. This "face-space" can continuously accommodate new faces and is built up and adapted from birth. Faces are positioned around a "central tendency" in this face-space according to their appearance: average faces cluster round the central

tendency, those that look similar are close together (e.g. a big nose, or round eyes), those that look dissimilar are far apart from each other, and distinct faces will be furthest from the central tendency. The main point is that each person's face-space is different, as what they consider to be a "typical" face will depend on their unique experience with faces.

Upon the assumption that primary faces in an infant's life are biological relatives (and thus of the same race), own-race faces will be judged to be more typical than other-race faces, and cluster round the central tendency. When the infant sees other-race faces, they will look very different to the faces that the infant is used to. Therefore, these other-race faces will be positioned far away from the centre of their personal face space. The main point is that the infant will have extensive experience with own-race faces, so it is easily able to individuate them, but the other-race faces are encountered rarely, so they are harder to individuate (Slater et al., 2010). A study that supports this model found that while people were better at individuating own-race faces, they were faster at classifying other-race faces as "other" than own-race faces as "own". The researchers concluded that this was probably because the other-race faces have fewer semantic representations, so they can be processed faster (Caldara, Rossion, Bovet, & Hauert, 2004).

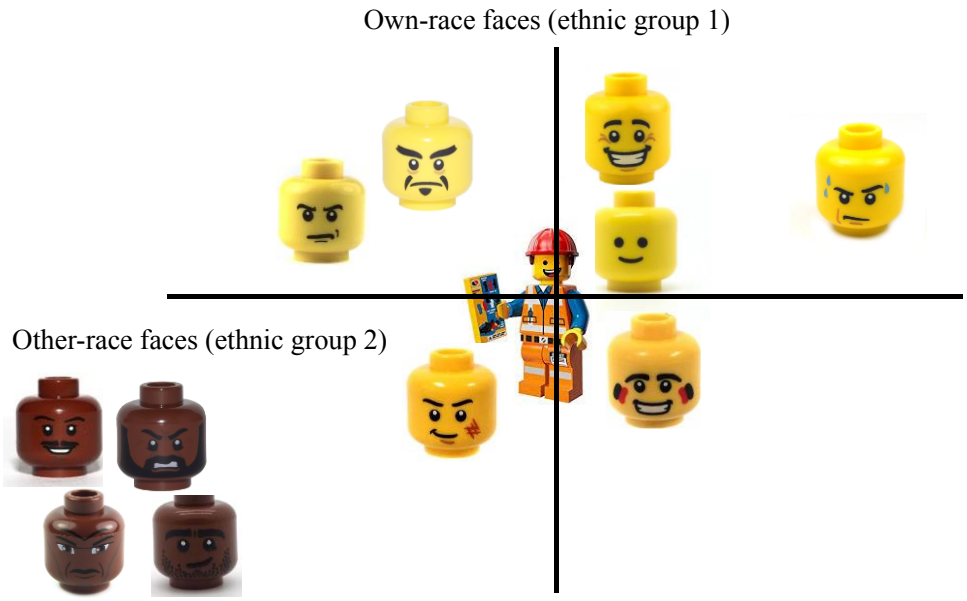


Fig. 3. An adaptation of Valentine's (1991) Exemplar-based model from his multi-dimensional face space framework.

(**legend**) Own-race faces cluster round the central tendency (spreading out according to visual similarity), while other-race faces cluster in a separate group, far away from the central tendency.

People sometimes have more experience with other-race faces than this model proposes, but the model can also accommodate this somewhat. The contact hypothesis suggests that the amount of contact with specific face types can moderate the ORE (e.g. Chiroro & Valentine, 1995; Wright, Boyd, & Tredoux, 2003), as contact increases expertise with other-race faces, making it easier to individuate them. Overall, it seems that having contact with other-race faces either makes it less likely that they will be positioned in a separate and distant cluster in face-space, or means that other-race face clusters will be less distinct.

An alternative account of the difficulty people have with recognising other-race faces is Sporer's (2001) In-group/Out-group Model (IOM), which suggests that face-race is one of the first things to be processed. If the face is own-race, it is processed configurally (to individuate it from other own-race faces). However, if it is other-race it is merely categorised as out-group and disregarded, so that processing that would help with individuation (distinguishing it from another face of that race) is compromised.

However, Sporer's (2001) model cannot account for the difficulty old people have with processing young faces, unless it could be extended to old people disregarding young faces in the same way that he proposed that other-race faces are. Old people were once young, so they would once have had considerable experience with other young faces, (this assumes that expertise endures, which might not be the case), but they find it more difficult to recognise young faces than faces of their own age group. One explanation for this effect (that Valentine's (1991) face-space can accommodate to some extent) is that as people age, young face types that once clustered around the central tendency drift further away and older face types drift inwards. This is because exposure to old faces increases as people age, so expertise with them also increases. Old faces thus become the "average" face types against which all other faces are judged (clustering around the central tendency). This makes them easier to individuate, at the expense of processing young faces. In turn, these become harder to individuate as they become increasingly distinct from the older faces against which they are judged. Face adaptation effects (Laurence & Hole, 2011; Laurence, Hole, & Hills, 2014) support this notion. They show that if a person looks at a "distorted" face for long enough, it will cease to look distorted. After normalising to the distorted face, subsequent "normal" faces appear distorted. This is because face-space may be calibrated in the visual system towards faces that are

commonly encountered. Thus, as a person ages the visual system may recalibrate as the encountered face-types change.

An alternative explanation is that people are more motivated to individuate socially-important faces, and that young faces lose their social-importance as people age. Indeed, socially-unimportant faces have been found to be treated as more unfamiliar than socially-important faces (Keyes & Zalicks, 2016), suggesting that motivation to individuate socially-important faces may be more important to face processing than physical appearance.

As stated above, contact has been found to improve recognition of other-race faces (e.g. Chiroro & Valentine, 1995; Wright et al., 2003), but it might be that rather than contact per se being key to face recognition, people are more motivated to individuate people with whom they spend time or who are socially-important to them. Indeed, Brigham and Malpass (1985) suggest that *frequent* contact is less important than *quality* contact. Harrison & Hole (2009) tested how contact affected the OAB using university students and trainee teachers with faces from 8-11 and 20-25 year olds, and found that the students showed OAB, but the trainee teachers who had extensive exposure to 8-11 year olds did not. While the researchers could not rule out perceptual expertise as an account of the effect, they proposed the notion that the trainee teachers were good at recognising the children because they were motivated to do so. As faces are socially-important, it makes sense that motivation to recognise is important to individuating between faces and learning new ones.

1.2. How are faces are learnt?

Little is known about the underlying processes involved in learning faces, although understanding this could help to understand how familiar and unfamiliar face processing differ. For instance, the type of exposure and/or time required for unfamiliar faces to appear familiar are issues which are not fully understood. However, some progress is now being made. For instance, Henderson, Williams, and Falk (2005) found that when the eye movements of participants were restricted as they were learning faces, they were less successful at recognising them later, suggesting that being able to move ones' eyes around the image of a face is important to learning it (see also Hills & Pake, 2013) rather than relying on peripheral vision.

Pilz et al., (2009) investigated whether movement assisted with face recognition, for faces that were learnt in an experiment. When participants learnt faces that moved, they performed better at test than when the faces were static, even if the faces were seen from a novel viewpoint at the test stage. This research suggests that seeing an unfamiliar face from multiple views increases its dynamic representation, making it easier to recognise later.

Dowsett, Sandford, & Burton (2015) found that the ability to account for within-face differences improved in a face-matching task when participants were given multiple different images of a target face (compared to just two), but performance was not at the levels expected for familiar face processing (e.g. Jenkins et al., 2011), suggesting that robust representations had not developed sufficiently during the experiment.

Longmore, Liu, & Young (2008) found that a single photograph did not provide enough information for a face to be recognised from another view, but showing

participants multiple views allowed the faces to be learnt sufficiently to be recognised from novel views. However, they conceded that even when showing participants multiple views, participants always performed best when viewing the original photograph in the test phase. This suggests that the representations of the faces in the experiment were not as robust as those of highly familiar faces. Thus while some experiments show a degree of face learning during an experiment (Kosaka et al., 2003; Henderson et al., 2005; Pilz et al., 2009; Dowsett et al,. 2015), Tong and Nakayama (1999) concluded that the development of robust representations takes far longer than can be investigated within a single experimental session.

While it is clear that some degree of learning is possible during an experiment, it is not clear whether this process is gradual or categorical. Dowsett, Sandford, & Burton (2015) found that faces were learnt incrementally during their experiment, and that matching performance improved as the number of different images of a target face was increased. Kosaka et al. (2003) repeatedly presented different images of unfamiliar faces, and found that activity in the bilateral posterior cingulate cortices also increased gradually, while it decreased in the right amygdala and left medial fusiform gyrus. However, Rossion et al. (2001) found that neurological activity in the right middle occipital gyrus, the right posterior fusiform gyrus, and the right inferotemporal cortex, changed abruptly as faces in an experiment became familiar, suggesting that face learning occurs suddenly. This is supported by Zimmermann and Eimer (2013), who found that the shift to view-independent face recognition occurs suddenly. However, the change in their study occurred after an experimental break, so learning could have evolved gradually during the break when the participants were not being tested. Therefore, it is still not known whether faces gradually become more and more robustly represented, or

whether they are categorised as unfamiliar until they reach a certain representation threshold, whereupon they are categorised as familiar.

1.2.1. Pupillometry as a measure of mental processing

Pupillometry is potentially a useful way to measure face learning because pupil size is not determined solely by ambient luminance, but can be influenced by mental processing, such as cognitive load. Research shows that the greater the mental workload, the larger the pupil size (Beatty, 1982; Jainta & Baccino, 2010; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; & see Ayres & Paas, 2012; Goldinger & Papesh, 2012; Murphy, Groeger, & Greene, 2016, for reviews). Pupil size has also been associated with affective processing, as pupils are larger when presented with emotional stimuli than with neutral stimuli (e.g. Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016). Pupillometry has also proved useful in indexing memory strength, as pupils have been shown to be larger when retrieving items associated with greater memory strength (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012), and they also appear to reflect the experience of recognition (Otero et al., 2011). Therefore, they may also reflect the strength of recognition evidence (Montefinese, Vinson, & Ambrosini, 2018).

Pupillometry also appears to be useful for measuring implicit memory, as pupillary changes occur in the absence of an overt response (van Rijn, Dalenberg, Borst, & Sprenger, 2012), and can even occur despite efforts to deceive. For instance, Heaver and Hutton (2011) found that pupil sizes were larger when looking at words that had been previously seen in a list, compared to new words. This was despite giving different instructions to participants, either to feign memory loss or to perform as accurately as

possible. Thus, pupil size reflected memory strength that was independent of the overt responses that participants gave.

Pupillometry has seldom been used in face recognition research. However, Goldinger, He, and Papesh (2009) have shown that pupil sizes were larger when looking at other-race faces than own-race faces. Considering the social importance of faces, and combining this with the findings that pupils respond to memory strength, it appears that pupillary changes could be a reliable measure of face recognition. It could also applicable to real-world applications such as eyewitness lineup procedures (review in Goldinger & Papesh, 2012).

1.2.2. Can pupillometry be used to measure Face learning?

Pupillometry has been shown to measure cognitive load (Piquado, Isaacowitz, & Wingfield, 2010; Chen & Epps, 2014; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014) and could be used to test whether faces are learnt gradually or abruptly. This is because cognitive load is lower when a task is easier than when it is difficult, and this is reflected in smaller pupil sizes (Jainta & Baccino, 2010). Therefore, face recognition should be easier when faces are familiar, and this should elicit smaller pupil sizes than when faces are unfamiliar.

Cognitive load theory (CLT) was designed to account for the effects of instructional design on cognitive load and learning (Sweller, 2010). Cognitive load is the notion that mental workload is affected by task demands (more difficult tasks will have higher loads than easier tasks) and distractions (caused by things like poor instructions), which increase cognitive load unnecessarily (see Moreno & Park, 2010; Sweller, 2010; Ayres & Paas, 2012; Murphy, Groeger, & Greene, 2016 for reviews). It is based on the

idea that there is a limited working-memory capacity and a long-lasting structure of long-term memory (Moreno & Park, 2010). To get memories from working-memory to long-term memory, it helps to pay attention to or rehearse them, but this is compromised when other information is also making demands.

Cognitive load theory thus applies well to face processing, as processing faces that have only been seen briefly before (and which are therefore unlikely to have been stored in long-term memory) probably places a greater cognitive burden on limited working-memory resources than processing familiar faces does. This is because briefly-seen faces have fragmented representations to be compared with, while familiar faces are represented by robust, dynamic and flexible representations, probably making them less effortful to process (see Hancock, Bruce, & Burton, 2000, for a review on mental representations).

Pupillometry is the measure of pupil sizes to make inferences about what is going on in the brain as people perform tasks. It is not known why pupils change size when there are fluctuations in cognitive processes. Indeed, there does not need to be an adaptive *reason* for the changes, which might merely be an epiphenomenon or a by-product of other adaptive attributes (see Gould & Lewontin, 1979 for a commentary on adaptationism). Pupillometry has been shown to measure fluctuations in cognitive processing that are inaccessible with other measures. For instance, pupillometry has been shown to measure memory strength in participants even when participants are asked to lie about their memory, thus responding independently from conscious decision responses (Heaver & Hutton, 2011).

Pupillary responses also have the advantage over neurological measures, which also measure implicit responses that are independent of decision responses, as the

technology is available to be used in applied settings. For instance, The EyeLink Duo (SR Research, n.d.) is a portable eye-tracking device that can be used with a laptop, and is able to measure accurate pupillary changes, even when participants are moving. Therefore, it is ideal to use with children. It can provide many pupillary scores including the mean, maximum and minimum pupil sizes for any trial or interest period (a period set by the experimenter), as well as eye-tracking measures, such as saccades (the abrupt movements from one point of focus to another) fixations (points of focus) and blinks. It can even account for pupils that appear elliptical due to the participant looking at something at the edge of the computer screen.

1.3. Aims of this thesis

The previous sections suggest that while face recognition has been widely researched, there remains relatively little understanding about how faces are learnt. Understanding that cognitive load is involved in learning, that pupillometry has been used to measure cognitive load, and knowing that familiar faces are easier to process than unfamiliar faces, we considered pupillometry to be an innovative way of measuring the demands made on mental resources by the task of face recognition. We considered that pupillometry could reflect cognitive processes during face learning, during familiar and unfamiliar face processing, and during eyewitness lineups. This is because pupillometry has been shown to measure fluctuations in cognitive load *as people are looking at objects*. Heaver and Hutton (2011) also showed that pupillary responses can be independent of conscious decision responses. Therefore, it may be possible to use pupillary responses to make inferences about the cognitive processes involved in tasks that are not always accessible via other means. This could be particularly useful in eyewitnesses, as identification responses are unreliable guides to recognition. While neurological markers

also probably achieve this, the technology required is not currently practical to administer. Thus, pupillometry has the potential to clarify theories and to be useful in practical settings.

The first half of the thesis (Chapters 2 & 3) aims to explore the theoretical accounts of pupillary changes while looking at different face types, starting with a broad investigation into cognitive load and face learning in Chapter 2. The second half of the thesis (Chapters 4-6) continues to evaluate theories, but focuses on forensic applications of pupillometry.

1.3.1. Summary of Chapter 2

The experiments presented in Chapter 2 were exploratory, investigating face learning and moderators to face learning, as well as various measures of face learning. We included all these measures, as no previous study has used pupil size to measure face learning. This resulted in multiple conditions in three separate experiments, the outcomes of which would inform the subsequent simpler experiments. Upon the assumptions that cognitive load is involved in face learning, and that pupillometry can measure cognitive load, we theorized that as faces became familiar, cognitive load would decrease, and so too would pupil size. Therefore, our first aim was to investigate whether pupillometry could reveal whether face learning occurred gradually or abruptly, by using experimentally-learnt (familiar) and experimentally-novel (unfamiliar) faces.

We also theorized (in line with CLT) that as cognitive load diminished, the learning outcomes would improve, resulting in greater levels of accuracy. Pupil sizes were thus compared to a more traditional explicit measure of learning, accuracy. We considered that if pupil sizes changed in similar ways to how accuracy changed, then we

could claim that they provided an indirect measure of face learning that supported the explicit measure of accuracy. However, if they changed *before* changes were seen in accuracy, it would suggest that pupil sizes provided a measure of implicit learning that occurred before conscious responses became more accurate. The advantage of pupillary markers of face recognition is that they are *independent* of explicit decisional processes that may be contaminated by the very act of making a conscious decision. Therefore, we combined traditional decision responses (accuracy) and pupil sizes as measures of face learning, in order to see whether face learning was gradual or abrupt.

We also measured reaction times (RTs), which have been used to measure cognitive processes in numerous experiments, and two other physiological responses: blinks (which have been associated with cognitive load) (Siegle, Ichikawa, & Steinhauer, 2008); and fixations (which have been associated with face learning) (Barton, Radcliffe, Cherkasova, Edelman, & Intriligator, 2006), to see whether they provided further information about the processes involved in face learning. Finally, we wanted to see whether the ORE (see Meissner & Brigham, 2001), the OAB (Anastasi & Rhodes, 2005), or gender (Wright & Sladden, 2003) affected face learning.

1.3.2. Cognitive Engagement

In Chapter 2 we found a reduction in pupil size that we attributed to a reduction in cognitive load as faces were learnt. Nevertheless, it is possible that this could have been attributed to diminishing engagement in the task, as previous research shows that pupil size can also be influenced by cognitively-engaging stimuli. For instance, pupils have been shown to be larger when participants are presented with emotional stimuli than with neutral stimuli (e.g. Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008a; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016),

and there are associations between large pupil sizes and physical attraction (Laeng & Falkenberg, 2007), goal-seeking (Mathôt, Siebold, Donk, & Vitu, 2015), and reward (Satterthwaite et al., 2007).

Therefore, the second theoretical construct discussed in this thesis is cognitive engagement. Cognitive engagement is currently a term used to describe motivation to invest in learning (Rotgans & Schmidt, 2011). However, in terms of this thesis, the construct is based upon the premise that salient objects (e.g. objects containing emotional content or social-importance) will be more engaging than non-salient objects (those with no emotional content or social-importance), and thus elicit larger pupil sizes than non-salient objects. Given that faces are socially important, it is likely that the degree to which a face engages a person will also affect pupil sizes, and this could out-weigh the changes that occur as a consequence of cognitive load. Based upon the research described above, it is therefore proposed that more cognitively engaging faces will result in larger pupil sizes than faces that are less engaging.

1.3.3. Summary of Chapter 3

Chapter 3 evaluated the relative accounts that cognitive load and cognitive engagement provide for changes in pupil size as people processed faces that we expected to differ in social importance.

Chapter 3 tested pupillary responses as participants looked at personally-familiar faces, unfamiliar faces and own faces. Faces that require greater cognitive load to process (unfamiliar faces) should also be less engaging than well-known faces. In other words, when cognitive load was greatest, cognitive engagement should be smallest. Therefore, if cognitive load places a greater burden on cognitive resources when processing

unfamiliar faces, then pupil sizes should be largest for unfamiliar faces; medium-sized for personally-familiar faces; and smallest for own faces, where the mental representation of the face should be the most robust (this assumption is discussed in detail in Chapter 3). However, if cognitive engagement best accounts for the pupillary changes, pupils should be largest when participants view their own face, due to a bias for own faces (Ninomiya, Onitsuka, Chen, Sato, & Tashiro, 1998; Kircher et al., 2001; Devue, Van der Stigchel, Brédart, & Theeuwes, 2009; Ramasubbu et al., 2011) and smallest when viewing the unfamiliar faces.

1.3.4. Does Pupillometry clarify the Cognitive Processes underlying Face Processing?

Previous research suggests that cognitive load and cognitive engagement can be separated (Moreno & Park, 2010), as cognitive load does not affect emotion processing (Berggren, Koster, & Derakshan, 2012), and it appears that cognitively engaging stimuli can be more distracting than objects that are associated with different degrees of cognitive load (Buetti & Lleras, 2016). This indicates that the responses to cognitively-engaging stimuli can override the demands of cognitive load.

While our experiments failed to conclude which theory best accounts for pupillary changes when processing faces, we tentatively considered that cognitive load diminished during the process of learning, and that unfamiliar faces were linked to larger pupil sizes, suggesting that they were more effortful to process. However, when faces differed in terms of social importance, cognitive engagement best accounted for the pupillary changes. The experiments in Chapters 2 and 3 set the scene in terms of exploring the theories behind accounts of pupillary changes, and provided us with some context before

testing pupillometry in forensic settings, which is the focus of the final sections of the thesis.

1.3.5. Can pupillometry be used as an index of face *recognition* that could be applied to forensic settings?

Research into the use of pupillometry for measuring face processing is very scarce. Goldinger, He, & Papesh (2009) found that pupil size reflected the Other Race Effect: pupil sizes were larger when looking at other-race faces than own-race faces. However, Goldinger and Papesh (2012) suggest in their review that pupillary responses could be a reliable measure of face recognition in eyewitness lineup procedures. Currently, the main reasons for expecting this to be the case are that pupils respond to cognitive processes including cognitive load, cognitive engagement, and memory strength (Heaver & Hutton, 2011), which are expected to be important in the processes involved with recognising faces in police lineups. In terms of cognitive load, this is because to the eyewitness the perpetrator's face would potentially be the only familiar face presented (at least in the first lineup presentation); in terms of cognitive engagement, the eyewitness should be motivated to recognise the perpetrator's face, which should also be the most socially-important of the faces they see; and in terms of memory strength, it should be the only face in the lineup that it is possible for them to remember. (The theory of memory strength will be discussed later in the thesis.)

1.4. Issues with Face Recognition in Forensic Settings.

This section will look at the ways in which face recognition research has informed our understanding of forensic issues, and tried to provide ways to reduce miscarriages of

justice. It will also evaluate whether pupillometry has any role in forensic procedures as a tool to measure face recognition, and whether the findings from Chapters 4, 5, and 6 can clarify the theories described above.

Research shows that unfamiliar face recognition is very difficult, yet this is what eyewitnesses are expected to do when making a lineup response. As discussed above, the ORE and OAB are well-known moderators of face recognition. For instance, in a task matching a travel-type document with an individual (Meissner, Susa, & Ross, 2013), it was found that participants matched own-race faces better than other-race faces. However, they were significantly more over-confident regarding their face matching accuracy for other-race faces than own-race faces. Thus, the ORE and OAB may have serious consequences in lineup procedures. The ORE in particular, has been investigated extensively in a forensic context (e.g. Wells & Olson, 2001; Wright & Stroud, 2002; Memon, Bartlett, Rose, & Gray, 2003; Havard & Memon, 2009; Havard, Memon, Laybourn, & Cunningham, 2012; Wylie, Bergt, Haby, Brank, & Bornstein, 2015), as other-race misidentification has been found to be 1.56 times higher than own-race misidentification, while own face identification was 1.4 times higher than other-race identification (Meissner & Brigham, 2001).

The internal state of the participant, e.g. stress (e.g. Valentine & Mesout, 2009; Rush et al., 2014; Attwood, Catling, Kwong, & Munafò, 2015) can also affect eyewitness identification. For instance, Steblay (1992) conducted a meta-analysis that suggested an effect of "weapon focus", where face recognition decreased in the presence of a weapon. It has also been shown that increased exposure to a face improves recognition accuracy and reduces false identifications, while increased delay (e.g. between seeing a “target” face and viewing a lineup of possible suspects) has the opposite effect (see MacLin,

MacLin, & Malpass, 2001 for a review). However, Read (1995) found that increased exposure time can *decrease* performance by increasing witnesses' readiness to make false identifications, as witnesses confused increased contextual familiarity (context) with increased perceptual familiarity for the face (recognition). Finally, Loftus, Schooler, Boone and Klein (1987) found that when people were stressed they overestimated the duration of events.

1.4.1. Improving Face Recognition in Forensic Settings.

Unfortunately, there is nothing that can be done about these "estimator variables" as they lie outside the control of the police, but awareness of them has helped police to understand the limitations of unfamiliar face recognition. However, procedures can be improved to minimise further obstacles to face processing accuracy.

For instance, double-blind techniques have improved identification reliability, e.g. Wells, Steblay, and Dysart (2015), as the person administering the lineup is not aware of who is the suspect, so is unable to influence the eyewitness. For instance, software such as VIPER (VIPER, n.d.) can be administered by someone who is not involved in the case. VIPER is a system that uses a large database of pre-recorded video clips of heads turning from centre, to the left, back to centre, to the right, and back to centre, filmed against a white background. When a suspect is found, they are filmed in the same way and matched to suitable "distractors" from the database, on the basis of the eyewitness's description of the perpetrator and the physical appearance of the suspect. This minimises bias (see Malpass, Tredoux, & McQuiston-Surrett, 2007, for a review of biased lineups) and makes lineups fairer. The eyewitness can even proceed unassisted, by following instructions on the screen, so that their responses are not influenced by a police officer. Also, as the distractors are taken from this database, they cannot be wrongfully convicted

(see Kemp, Pike, and Brace, 2001, for a commentary). Thus, VIPER has dramatically improved lineup procedures in the UK.

The method of presenting faces to the eyewitness is also important. There are currently variations of two approved methods: simultaneous and sequential. Simultaneous lineups show photographic images of all the faces at the same time. They are associated with more correct identifications than sequential lineups, but also more misidentifications, as they encourage people to make relative judgements. This means that people compare lineup faces for the closest match to their memory of the suspect, and pick the face that is the best fit to their memory of the perpetrator (Flowe & Cottrell, 2011). Most law enforcement agencies in the US use the simultaneous system containing six faces (Seale-Carlisle & Mickes, 2016).

Sequential lineups show images (either photographs or videos) one at a time. Research shows that people make more correct identifications in simultaneous lineups, but at the expense of also making more misidentifications. This is because sequential lineups encourage absolute judgments, meaning that the witness compares the face being assessed to their memory of the perpetrator (Cutler and Penrod, 1988; Lindsay & Wells 1985; Sporer 1993). However, both procedures only produce about 25% correct identifications overall (Wells, Steblay, & Dysart, 2015).

In the UK, a hybrid sequential system (containing nine faces) is used, where there are two presentations of a sequential display (Seale-Carlisle & Mickes, 2016). There is a similar procedure in the US, but in this system eyewitnesses have an *option* to see the second lineup presentation. This was tested by Steblay, Dietrich, Ryan, Raczynski & James (2011). They found that people made more identifications in the second lineup than the first, and that participants who *chose* to have two presentations were less accurate

than those who did not. These participants also performed worse in the second lineup presentation than they had in the first. Therefore, hybrid designs are unhelpful in improving eyewitness performance.

1.4.2. Measuring Recognition in Eyewitnesses

It is known that eyewitnesses' identification responses are inaccurate, so research has tried to design ways to test the eyewitness's credibility (see The Turnbull Guidelines, 1977, available at CPS (n.d.) for eyewitness credibility guidelines) by measuring the likelihood that an identification is reliable.

Bindemann, Brown, Koyas, & Russ (2012) attempted to predict identification accuracy by comparing scores on established face recognition tests with lineup responses. They used the 1-in-10 face recognition test (Bruce et al., 1999), which presents one target face and ten "distractor" faces. The participant has to decide whether the target face also appears among the distractor faces. The main points are that all the faces are shown at the same time, so participants can make direct comparisons, and that the main target image is different from the image of the target in the distractor display (in target-present displays). Bindemann et al. found that the task provided a good index of eyewitness reliability for participants who made an identification (a correct identification or a misidentification), but not for those who made no identification (no identification or a correct rejection).

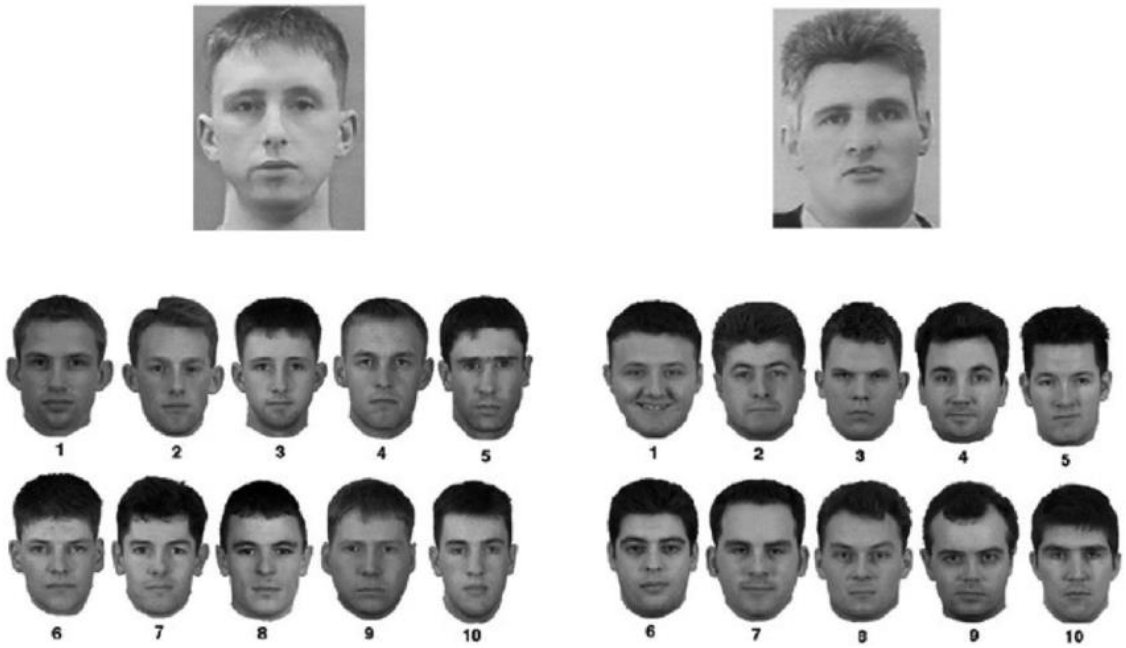


Fig. 4. Two examples of the 1-in-10 task (Bruce et al., 1999): left-hand side: a target-present array; right-hand side: a target-absent array.

Confidence has also been studied. It is usually measured using confidence rating scales (e.g. 1 = not confident, 5 = very confident). Not all research has found confidence to be reliable as a measure of identification performance (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988), as confidence ratings can be influenced by post-identification feedback. These effects are worse when the confidence rating is made after a delay rather than immediately after viewing the lineup (Wells & Bradfield, 1998). Post identification feedback is when an eyewitness is told whether or not their response is correct (this is particularly problematic when police give feedback, as they have no way of knowing whether or not the suspect is actually the perpetrator). However, confidence ratings can be useful when a witness makes an identification (Sauerland & Sporer, 2009; Sauer, Brewer, Zweck, and Weber (2009), providing that they are recorded immediately after an identification is made and before any feedback is given. Therefore, it is recommended

that confidence is recorded immediately after a lineup (Wells et al., 1998; National Academy of Sciences, 2014) rather than at a later date.

An alternative measure to confidence is the remember-know (RK) paradigm. This is a self-rating measure of a participant's belief in their own memory strength. "Remember" (R) is chosen when participants believe they remember the to-be-remembered item well, and "Know" (K) is chosen when they think they know the answer without actually remembering the item. Tulving (1985) introduced it in order to measure states of awareness that were considered to underlie memory retrieval. Thus, it is probably appropriate to use with eyewitnesses, as they are asked to give memory-based evidence. As such, versions that include a "Guess" ("G") response option are more useful, as not all eyewitnesses can claim to have any memory of the perpetrator.

One option is to use neurological markers of face recognition. For instance, Lefebvre, Marchand, Smith, and Connolly (2007) found that participants who made correct identifications showed an increased P300 response to the target compared to distractors. The P300 was also significantly larger in participants who correctly identified the target than in those who misidentified the target. The advantage of neurological and physiological markers of face recognition is that they are *independent* of the explicit decisional processes involved in making an identification. Measures of witness confidence, self-ratings of memory, or measures of generalised face recognition ability do not fulfil this criterion, as they are alternative explicit measures of recognition that may be contaminated by the conscious decision processes involved in responding in the first place. For instance, self-ratings scales may be associated with motivation or self-perception (Kassin, Rigby, Castillo, 1991), or self-concept (Kröner & Biermann, 2007).

However, the technologies currently required to measure neurological markers of face recognition are too impractical to use in forensic settings. Pupillometry has the potential to help, as it fulfils the criterion of measuring a physiological response that is outside of the witness' conscious control (more likely to be independent of their overt decision), and it is practical to use. Therefore, pupillometry could be a useful supplementary measure of eyewitness identification performance.

1.4.3. Summary of Chapters 4 & 5

Chapter 4 was the first of three that explored the role of pupillometry in forensic settings, by investigating pupillary responses to a target face in a lineup, something that had not been done before (to our knowledge). The chapter tested participants' performance with a hybrid lineup (with two sequential presentations) after having seen a video of a mock crime, and recorded their pupil sizes as they did so. We compared pupillary changes in the target-present condition with those in a target-absent condition. Finally, we measured participants' subjective assessment of their memory (using the RKG paradigm) to see whether they had any insight into their performance, as previous research suggests that pupil sizes can be affected by memory strength. It seemed plausible that memory strength would account for pupillary changes in a lineup more than cognitive load or cognitive engagement, so we evaluated pupillary responses in light of all three explanations. We found that pupillometry was a useful measure of implicit recognition strength, and that pupillary responses were indeed independent of explicit identification responses.

In Chapter 5, we extended the research that we had conducted in Chapter 4, by using methods more in line with those used by UK police, to see whether pupillometry

was viable in the procedures that they currently use. We found that pupillometry was equally useful in UK Police style lineup procedures.

1.4.4. Summary of Chapter 6

In Chapter 6, we extended our previous research in three ways: by asking participants to identify a person they had just met (rather than using a video); by testing the viability of using pupillometry on anxious people; and by doing so in a field experiment. We recruited people at the British Science Festival (2017), as they came off a scary ride on Brighton Pier. They were approached by a female researcher who asked them to complete a questionnaire designed to measure their level of self-reported anxiety. After this, they viewed a UK style hybrid lineup, and were asked to decide whether or not the face of the researcher they had seen at the ride was present in the lineup and if so, to identify her. As in the previous chapters, pupillometry successfully measured implicit memory strength as participants looked at the lineup. The study was designed to test the viability of portable eye-trackers as a way to collect eyewitness data in real-world settings, and was used as a starting point for research into the use of pupillometry in forensic settings.

1.5. Summary of the aims of the thesis.

Our primary aim was to investigate whether pupillometry could serve as a marker of the processes involved in face recognition. In particular, we aimed to investigate familiar and unfamiliar face processing, face learning, and face recognition in forensic settings. We also evaluated the pupillary results in relation to explanations in terms of cognitive load, cognitive engagement and memory strength.

References

Allport, G. W., & Postman, L. (1947). *The psychology of rumor*. New York: Russell & Russell.

Anaki, D., Boyd, J., & Moscovitch, M. (2007). Temporal integration in face perception: Evidence of configural processing of temporally separated face parts. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 1–19. http://dx.doi.org.ezproxy.sussex.ac.uk/10.1037/0096-1523.33.1.1

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043–1047.

Attwood, A. S., Catling, J. C., Kwong, A. S. F., & Munafò, M. R. (2015). Effects of 7.5% carbon dioxide (CO2) inhalation and ethnicity on face memory. *Physiology & Behavior*, *147*, 97–101. https://doi.org/10.1016/j.physbeh.2015.04.027

Avidan, G., Tanzer, M., & Behrmann, M. (2011). Impaired holistic processing in congenital prosopagnosia. *Neuropsychologia*, *49*(9), 2541–2552. https://doi.org/10.1016/j.neuropsychologia.2011.05.002

Ayres, P., & Paas, F. (2012). Cognitive Load Theory: New directions and challenges: Cognitive load theory: new directions. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Balas, B. (2012). Bayesian face recognition and perceptual narrowing in face-space: Bayesian face recognition and perceptual narrowing. *Developmental Science*, *15*(4), 579–588. https://doi.org/10.1111/j.1467-7687.2012.01154.x

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*, *35*(8), 1089–1105. https://doi.org/10.1068/p5547

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276-292.

Berggren, N., Koster, E. H. W., & Derakshan, N. (2012). The effect of cognitive load in emotional attention and trait anxiety: An eye movement study. *Journal of Cognitive Psychology*, *24*(1), 79–91. https://doi.org/10.1080/20445911.2011.618450

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brédart, S. (2003). Recognising the usual orientation of one's own face: The role of asymmetrically located details. *Perception*, *32*(7), 805–811. https://doi.org/10.1068/p3354

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364.

Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of faces of own-and other-race persons. *Journal of Social Issues, 41*(3), 139-155.

British Science Festival (2017), retrieved 8th February 2018, from https://www.britishscienceassociation.org/british-science-festival

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology*, *53*(12), 1823–1835. https://doi.org/10.1111/psyp.12770

Bruce, V., Burton, A. M., Craw, I., Valentine, T., Rolls, E. T., & Ellis, H. D. (1992). Modelling face recognition [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 121–128.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339–360.

Buetti, S., & Lleras, A. (2016). Distractibility is a function of engagement, not task difficulty: Evidence from a new oculomotor capture paradigm. *Journal of*

Experimental Psychology: General, *145*(10), 1382–1405. https://doi.org/10.1037/xge0000213

Bushnell, I.W.R. (1998). The origins of face perception. *The Development of Sensory, Motor and Cognitive Capacities in Early Infancy*, Simion F, Butterworth G (eds). Psychology Press: Hove, East Sussex; 69–86.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Busey, T. A., Brady, N. P., & Cutting, J. E. (1990). Compensation is unnecessary for the perception of faces in slanted pictures. *Perception & Psychophysics*, *48*(1), 1–11.

Cabeza, R., & Kato, T. (2000). Features are also important: Contributions of featural and configural processing to face recognition. *Psychological Science, 11*(5), 429-433.

Caldara, R., Rossion, B., Bovet, P., & Hauert, C.-A. (2004). Event-related potentials and time course of the 'other-race' face classification advantage. *Neuroreport*, *15*(5), 905–910.

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 527–551. http://dx.doi.org.ezproxy.sussex.ac.uk/10.1037/0096-1523.26.2.527

Carbon, C.-C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801–806. https://doi.org/10.1068/p5789

Chatterjee, A., Thomas, A., Smith, S. E., & Aguirre, G. K. (2009). The neural response to facial attractiveness. *Neuropsychology*, *23*(2), 135–143. https://doi.org/10.1037/a0014430

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Cheung, O. S., Richler, J. J., Palmeri, T. J., & Gauthier, I. (2008). Revisiting the role of spatial frequencies in the holistic processing of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1327–1336. https://doi.org/10.1037/a0011752

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 879–894. https://doi.org/10.1080/14640749508401421

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116. https://doi.org/10.1080/09541440340000439

Collishaw, S. M., & Hole, G. J. (2000). Featural and configurational processes in the recognition of faces of different familiarity. *Perception*, *29*(8), 893–909. https://doi.org/10.1068/p2949

Cutler, B. L., & Penrod, S. D. (1988). Improving the reliability of eyewitness identification: Lineup construction and presentation. *Journal of Applied Psychology, 73*(2), 281-290.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*(1), 41-56.

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, *10*(6), 393–396.

Dalrymple, K. A., Fletcher, K., Corrow, S., das Nair, R., Barton, J. J. S., Yonas, A., & Duchaine, B. (2014). "A room full of strangers every day": The psychosocial impact of developmental prosopagnosia on children and their families. *Journal of Psychosomatic Research*, *77*(2), 144–150. https://doi.org/10.1016/j.jpsychores.2014.06.001

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*(6), 687-706.

DeGutis, J., DeNicola, C., Zink, T., McGlinchey, R., & Milberg, W. (2011). Training with own-race faces can improve processing of other-race faces: Evidence from developmental prosopagnosia. *Neuropsychologia*, *49*(9), 2505–2513. https://doi.org/10.1016/j.neuropsychologia.2011.04.031

Devue, C., Van der Stigchel, S., Brédart, S., & Theeuwes, J. (2009). You do not find your own face faster; you just look at it longer. *Cognition*, *111*(1), 114–122. https://doi.org/10.1016/j.cognition.2009.01.003

Dowsett, A. J., Sandford, A., & Burton, A. M. (2015). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *The Quarterly Journal of Experimental Psychology*, 1–10. https://doi.org/10.1080/17470218.2015.1017513

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431–439.

Fink, B., Neave, N., Manning, J. T., & Grammer, K. (2006). Facial symmetry and judgements of attractiveness, health and personality. *Personality and Individual Differences*, *41*(3), 491–499. https://doi.org/10.1016/j.paid.2006.01.017

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239-260.

Flowe, H., & Cottrell, G. W. (2011). An examination of simultaneous lineup identification decision processes using eye tracking. *Applied Cognitive Psychology*, *25*(3), 443–451. https://doi.org/10.1002/acp.1711

Frowd, C. D., Carson, D., Ness, H., McQuiston-Surrett, D., Richardson, J., Baldwin, H., & Hancock, P. (2005). Contemporary composite techniques: The impact of a forensically-relevant target delay. *Legal and Criminological Psychology*, *10*(1), 63–81. https://doi.org/10.1348/135532504X15358

George, P. A., & Hole, G. J. (1995). Factors influencing the accuracy of age estimates of unfamiliar faces. *PERCEPTION-LONDON-*, *24*, 1059–1059.

Goffaux, V., & Rossion, B. (2006). Faces are" spatial"–holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 1023-1039.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science, 21*(2), 90-95. https://doi.org/10.1177/0963721412436811

Goldstein, A. G., & Mackenberg, E. J. (1966). Recognition of human faces from isolated facial features: A developmental study. *Psychonomic Science, 6*(4), 149-150.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B, 205*(1161), 581-598.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Harrison, V., & Hole, G. J. (2009). Evidence for a contact-based explanation of the own-age bias in face recognition. *Psychonomic Bulletin & Review*, *16*(2), 264–269. https://doi.org/10.3758/PBR.16.2.264

Havard, C., & Memon, A. (2009). The influence of face age on identification from a video line-up: A comparison between older and younger adults. *Memory*, *17*(8), 847–859. https://doi.org/10.1080/09658210903277318

Havard, C., Memon, A., Laybourn, P., & Cunningham, C. (2012). Own-age bias in video lineups: a comparison between children and adults. *Psychology, Crime & Law*, *18*(10), 929–944. https://doi.org/10.1080/1068316X.2011.598156

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, *33*(1), 98–106.

Hills, P. J., & Pake, J. M. (2013). Eye-tracking the own-race bias in face recognition: Revealing the perceptual and socio-cognitive mechanisms. *Cognition, 129*(3), 586-597.

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception*, *23*(1), 65–74.

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221–1240. https://doi.org/10.1068/p3252

The Innocence Project (n.d.), retrieved 18th May, 2018, from https://www.innocenceproject.org/causes/eyewitness-misidentification/

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671-1683. https://doi.org/10.1098/rstb.2010.0379

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnston, P. R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596. https://doi.org/10.1080/09658210902976969

Kassin, S. M., Rigby, S., & Castillo, S. R. (1991). The accuracy-confidence correlation in eyewitness testimony: Limits and extensions of the retrospective self-awareness effect. *Journal of Personality and Social Psychology*, *61*, 698–707.

Kemp, R. I., Pike, G. E., & Brace, N. A. (2001). Video-based identification procedures: Combining best practice and practical requirements when designing identification systems. *Psychology, Public Policy, and Law*, *7*(4), 802–807. https://doi.org/10.1037//1076-8971.7.4.802

Keyes, H., & Zalicks, C. (2016). Socially important faces are processed preferentially to other familiar and unfamiliar faces in a priming task across a range of viewpoints. *PLOS ONE*, *11*(5), e0156350. https://doi.org/10.1371/journal.pone.0156350

Kircher, T. T. J., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P. J., Bullmore, E. T., … David, A. S. (2001). Recognizing one's own face. *Cognition*, *78*(1), B1–B15. https://doi.org/10.1016/S0010-0277(00)00104-9

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition*, *4*(3), 265–273.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2013). Effects of aging on face identification and holistic face processing. *Vision Research*, *88*, 38–46. https://doi.org/10.1016/j.visres.2013.06.003

Kosaka, H., Omori, M., Iidaka, T., Murata, T., Shimoyama, T., Okada, T., … Wada, Y. (2003). Neural substrates participating in acquisition of facial familiarity: an fMRI study. *NeuroImage*, *20*(3), 1734–1742. https://doi.org/10.1016/S1053-8119(03)00447-6

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept—Towards a model of response confidence. *Intelligence*, *35*(6), 580-590.

Laeng, B., & Falkenberg, L. (2007). Women's pupillary responses to sexually significant others during the hormonal cycle. *Hormones and Behavior*, *52*(4), 520–530. https://doi.org/10.1016/j.yhbeh.2007.07.013

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, *12*(4), 259–272. https://doi.org/10.1207/S15326969ECO1204_01

Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition*, *10*(8), 897–912. https://doi.org/10.1080/13506280344000149

Laurence, S., & Hole, G. (2011). The effect of familiarity on face adaptation. *Perception*, *40*(4), 450–463. https://doi.org/10.1068/p6774

Laurence, S., Hole, G. J., & Hills, P. J. (2014). Lecturers' faces fatigue their students: Face identity aftereffects for dynamic and static faces. *Visual Cognition*, *22*(8), 1072–1083. https://doi.org/10.1080/13506285.2014.950364

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, *44*(6), 894–904. https://doi.org/10.1111/j.1469-8986.2007.00566.x

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564.

Loftus, E. F., Schooler, J. W., Boone, S. M., & Kline, D. (1987). Time went by so slowly: Overestimation of event duration by males and females. *Applied Cognitive Psychology, 1,* 3–13. doi: 10.1002/acp.2350010103

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77–100. http://dx.doi.org.ezproxy.sussex.ac.uk/10.1037/0096-1523.34.1.77

MacLin, O. H., MacLin, M. K., & Malpass, R. S. (2001). Race, arousal, attention, exposure and delay: An examination of factors moderating face recognition. *Psychology, Public Policy, and Law*, *7*(1), 134–152. https://doi.org/10.1037//1076-8971.7.1.134

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 155–178). Lawrence Erlbaum Mahwah, NJ.

Mathôt, S., Siebold, A., Donk, M., & Vitu, F. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, *144*(3), 513–521. https://doi.org/10.1037/a0039168

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175–1184.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364–372. https://doi.org/10.1037/a0013464

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology, 64*(8), 1473-1483.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037//1076-8971.7.1.3

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own- and other-race faces. *Visual Cognition*, *21*(9–10), 1287–1305. https://doi.org/10.1080/13506285.2013.832451

Memon, A., Bartlett, J., Rose, R., & Gray, C. (2003). The aging eyewitness: Effects of age on face, delay, and source-memory ability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*(6), 338–345. https://doi.org/10.1093/geronb/58.6.P338

Michel, C., Rossion, B., Han, J., Chung, C.-S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608–615.

Mondloch, C. J., Lewis, T. L., Budreau, D. R., Maurer, D., Dannemiller, J. L., Stephens, B. R., & Kleiner-Gathercoal, K. A. (1999). Face perception during early infancy. *Psychological Science*, *10*(5), 419–422.

Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: the pupil's point of view. *Biological psychology*, *135*, 159-169.

Moreno, R., & Park, B. (2010). Cognitive Load Theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 9–28). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.003

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic bulletin & review, 23*(5), 1316-1340. https://doi.org/10.3758/s13423-015-0982-5

Murray, D. C. (2015). Notes to self: the visual culture of selfies in the age of social media. *Consumption Markets & Culture*, *18*(6), 490–516. https://doi.org/10.1080/10253866.2015.1052967

National Academy of Sciences. (2014). Using eyewitness identifications: New report urges caution. *ScienceDaily*. Retrieved May 1, 2018 from www.sciencedaily.com/releases/2014/10/141002123735.htm

Ninomiya, H., Onitsuka, T., Chen, C.-H., Sato, E., & Tashiro, N. (1998). P300 in response to the subject's own face. *Psychiatry and Clinical Neurosciences*, *52*(5), 519–522. https://doi.org/10.1046/j.1440-1819.1998.00445.x

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory: Pupil size and recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

O'toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, *22*(2), 208–224.

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pilz, K. S., Bülthoff, H. H., & Vuong, Q. C. (2009). Learning influences the encoding of static and dynamic faces and their recognition across different spatial frequencies. *Visual Cognition*, *17*(5), 716–735. https://doi.org/10.1080/13506280802340588

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Prehn, K., Heekeren, H. R., & van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Ramasubbu, R., Masalovich, S., Gaxiola, I., Peltier, S., Holtzheimer, P. E., Heim, C., … Mayberg, H. S. (2011). Differential neural activity and connectivity for processing one's own face: A preliminary report. *Psychiatry Research: Neuroimaging*, *194*(2), 130–140. https://doi.org/10.1016/j.pscychresns.2011.07.002

Read, J. D. (1995). The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology*, *9*(2), 91-121.

Rhodes, G. (1985). Lateralized processes in face recognition. *British journal of Psychology*, *76*(2), 249-271.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision Research*, *49*(23), 2856–2861. https://doi.org/10.1016/j.visres.2009.08.025

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, Mechanisms, and Measures of Holistic Processing. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00553

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: a PET study of face categorical perception. *Journal of Cognitive Neuroscience*, *13*(7), 1019–1034.

Rotgans, J. I., & Schmidt, H. G. (2011). Cognitive engagement in the problem-based learning classroom. *Advances in Health Sciences Education*, *16*(4), 465–479. https://doi.org/10.1007/s10459-011-9272-9

Rush, E. B., Quas, J. A., Yim, I. S., Nikolayev, M., Clark, S. E., & Larson, R. P. (2014). Stress, interviewer support, and children's eyewitness identification accuracy. *Child Development*, *85*(3), 1292–1305. https://doi.org/10.1111/cdev.12177

Sandford, A., & Burton, A. M. (2014). Tolerance for distorted faces: Challenges to a configural processing account of familiar face recognition. *Cognition*, *132*(3), 262–268. https://doi.org/10.1016/j.cognition.2014.04.005

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage*, *37*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*(4), 337-347.

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*, *3*(9), 160300. https://doi.org/10.1098/rsos.160300

Senft, T. M., & Baym, N. K. (2015). Selfies introduction: What does the selfie say? Investigating a global phenomenon. *International Journal of Communication*, *9*, 19, 1588-1606.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*(5), 679–687. https://doi.org/10.1111/j.1469-8986.2008.00681.x

Slater, A., Quinn, P. C., Kelly, D. J., Lee, K., Longmore, C. A., McDonald, P. R., & Pascalis, O. (2010). The Shaping of the Face Space in Early Infancy: Becoming a Native Face Processor: Becoming a Native Face Processor. *Child Development Perspectives*, *4*(3), 205–211. https://doi.org/10.1111/j.1750-8606.2010.00147.x

Smith, E. E., & Nielsen, G. D. (1970). Representations and retrieval processes in short-term memory: Recognition and recall of faces. *Journal of Experimental Psychology, 85*(3), 397-405.

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78(*1), 22-33.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36-97.

SR Research (n.d.), retrieved, April 23rd, 2018, from https://www.sr-research.com/products/eyelink-1000-plus/

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262–274. https://doi.org/10.1007/s10979-010-9236-2

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 29–47). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Tacikowski, P., & Nowicka, A. (2010). Allocation of attention to self-name and self-face: An ERP study. *Biological Psychology*, *84*(2), 318–324. https://doi.org/10.1016/j.biopsycho.2010.03.009

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*(2), 225–245.

Tong, F., & Nakayama, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1016–1035. http://dx.doi.org/10.1037/0096-1523.25.4.1016

Tulving, E., & Murray, D. (1985). Elements of episodic memory. *Canadian Psychology, 26*(3), 235-238.

The Turnbull Guidelines (1977), retrieved 18th May, 2018, from https://www.cps.gov.uk/legal-guidance/identification

Tversky, A., & Krantz, D. H. (1969). Similarity of schematic faces: A test of interdimensional additivity. *Perception & Psychophysics, 5*(2), 124-128.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T. & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica*, *61*(3), 259–273.

Valentine, T, & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, *23*(2), 151–161. https://doi.org/10.1002/acp.1463

Van Belle, G., De Graef, P., Verfaillie, K., Busigny, T., & Rossion, B. (2010). Whole not hole: Expert face recognition requires holistic perception. *Neuropsychologia*, *48*(9), 2620–2629. https://doi.org/10.1016/j.neuropsychologia.2010.04.034

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS ONE*, *7*(12), e51134. https://doi.org/10.1371/journal.pone.0051134

VIPER (n.d.), retrieved January 16th, 2018, from http://www.viper.police.uk

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360-376.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, *22*(6), 603-647.

Wells, G. L., & Olson, E. A. (2001). The other-race effect in eyewitness identification: What do we do about it? *Psychology, Public Policy, and Law*, *7*(1), 230–246. https://doi.org/10.1037//1076-8971.7.1.230

Wells, G. L., & Olson, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, *39*(1), 1–14. https://doi.org/10.1037/lhb0000096

Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2003). Inter-racial contact and the own-race bias for face recognition in South Africa and England. *Applied Cognitive Psychology*, *17*(3), 365–373. https://doi.org/10.1002/acp.898

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica*, *114*(1), 101–114. https://doi.org/10.1016/S0001-6918(03)00052-0

Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: people are better with their own age. *Law and Human Behavior*, *26*(6), 641.

Wylie, L. E., Bergt, S., Haby, J., Brank, E. M., & Bornstein, B. H. (2015). Age and lineup type differences in the own-race bias. *Psychology, Crime & Law*, *21*(5), 490–506. https://doi.org/10.1080/1068316X.2014.989173

Xiao, N. G., Quinn, P. C., Ge, L., & Lee, K. (2012). Rigid facial motion influences featural, but not holistic, face processing. *Vision Research*, *57*, 26–34. https://doi.org/10.1016/j.visres.2012.01.015

Yardley, L., McDermott, L., Pisarski, S., Duchaine, B., & Nakayama, K. (2008). Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of Psychosomatic Research*, *65*(5), 445–451. https://doi.org/10.1016/j.jpsychores.2008.03.013

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737–746.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Zimmermann, F. G. S., & Eimer, M. (2013). Face learning and the emergence of view-independent face recognition: An event-related brain potential study. *Neuropsychologia*, *51*(7), 1320–1329. https://doi.org/10.1016/j.neuropsychologia.2013.03.028

CHAPTER 2. SLOW AND STEADY WINS THE FACE: MEASURING FACE LEARNING WITH PUPILLOMETRY.

Abstract

It is understood that familiar faces are processed differently to unfamiliar faces. However, it is not known how this is associated with differences in cognitive load. Little is also known about the ways in which faces are learnt, or whether face learning is associated with a reduction in cognitive load. This study compared explicit (accuracy and speed of familiarity judgements to faces) and implicit (physiological) measures of face processing. Our focus was on pupillary responses, which have been associated with cognitive load. We investigated how they were affected by familiar and unfamiliar face processing, how they were affected by repeated presentations of novel images of faces, and whether they were related to improvements in accuracy. We found that processing familiar faces was associated with smaller pupils, suggesting that they produce less cognitive load than unfamiliar faces. Pupil size and accuracy showed similar patterns over time, supporting the idea that faces are learnt gradually. It appears that face learning is associated with a diminishing reduction in cognitive load, which is mediated by differences in the age, race or gender between the participant and the face seen.

Face recognition research suggests that recognising familiar faces is relatively easy, while recognising faces that have only been seen once or twice before is considerably harder. For example, Bruce, Henderson, Newman and Burton (2001) found that participants matched CCTV images of unfamiliar faces successfully only 70% of the time, while they successfully matched images of familiar faces 92% of the time. It has been proposed that familiar and unfamiliar face recognition use different processes (Hancock, Bruce, & Burton 2000). Burton, Jenkins, and Schweinberger (2011) suggest that unfamiliar face processing is based on inflexible pictorial codes that can be used to determine whether or not a particular image has been seen before. However, familiar face processing is based on abstract structural codes: when a familiar face is seen, its characteristics can be matched to its stored representation, even when the face is seen under novel conditions. Zimmermann and Eimer's research (2013) supports this distinction, as they found that unfamiliar face recognition is view-dependent, while familiar face recognition is possible across different views. This explains why familiar face recognition is good even when image quality is poor, and why unfamiliar face recognition is poor even when image quality is good.

Research has also investigated the Own Age Bias (OAB), which indicates that own-age faces are easier to recognise or learn than faces of a markedly different age (see Rhodes & Anastasi, 2012 for a review). Anastasi & Rhodes (2005) tested children (aged 5-8) and older people (aged 55-89) and found that both age groups recognised own-age faces more accurately than other-age faces. Inferior recognition may be partly due to a lack of contact with other face types, as while old people were once young and would once have had frequent contact with other young people, the frequency of this contact may have diminished as they aged. However, motivation to individuate out-group members improves people's ability to recognise faces from a different age group to their

own (e.g. Harrison & Hole; 2009; Proietti, Pisacane, & Macchi Cassia, 2013), suggesting that social importance can moderate recognition (Keyes & Zalicks, 2016).

While we know that familiar and unfamiliar faces are processed somewhat differently, relatively little is known about the underlying processes of face learning, although some progress is now being made. For example, Longmore, Liu, & Young (2008) found that unfamiliar faces remained poorly recognised after participants viewed multiple exposures of the same photograph, but could be recognised more easily after being seen from multiple views. This suggests that seeing multiple views of a face either increases the amount of information about that face, making it easier to recognise later, or allows dynamic representations of faces to be made, which are more flexible for future recognition.

However, the time-course of the transition from being "unfamiliar" to "familiar" remains uncertain, which may be due in part to different definitions of face learning, different tasks and different understandings of what constitutes familiarity. For instance, Tong and Nakayama (1999) found that developing robust representations of faces involves protracted experience with the faces, much longer than can be investigated within the fairly short sessions typical of most experimental studies. However, Pilz, Bülthoff, & Vuong (2009) found that face learning can occur within a single experimental session, and that it occurred gradually (see also Kosaka et al., 2003). This suggests that the two experiments perhaps had different definitions of familiarity.

Tong and Nakamaya (1999) used static black and white images of participants' own faces and experimentally-learnt faces in a task that required participants to select either the only ‘own face’ image from five distractors that were displayed at the same time, or the only ‘stranger’ from several distractors. They found that reaction times when

looking at the stranger did decrease during the early part of the experiment. However, RTs were always faster for own-face images, even when the images of the 'stranger faces' became more familiar during the experiment. Numerous studies show that there is a 'self bias' in that own faces attract more attention and are responded to faster than other faces (e.g. Devue & Brédart, 2008; Devue, Van der Stigchel, Brédart, & Theeuwes, 2009; Tacikowski & Nowicka, 2010), so the ways in which 'own face' images are processed may not be comparable to the processing of other highly familiar faces.

Pilz et al. (2009) used a same/different task, where participants saw a prime followed by a target image (all images were in colour). This was either the same person as the prime or a different person. The researchers found that participants responded faster to faces that were learnt in motion. Therefore, while the threshold of 'familiarity' and the procedures were different between the two experiments, making them difficult to compare, they both found that reaction times were faster when looking at faces that were more robustly represented.

In contrast, Rossion, Schiltz, Robaye, Pirenne, and Crommelinck (2001) tested participants on experimentally-learnt black and white images of faces that were morphed with unfamiliar faces to different degrees (0% - 100%). When participants were presented with faces along this continuum of familiarity and asked whether the face was *familiar* or not, familiarity ("familiar"/"unfamiliar") decisions changed between the 40%-60% morphs. Most importantly neurological activity also changed abruptly between the 40%-60% morphs when participants were asked to categorise the faces *according to their gender*. This suggests that faces are categorised as either familiar or unfamiliar (rather than evaluated along a continuum of familiarity) even when participants are not asked to distinguish them according to familiarity.

Finally, Zimmermann and Eimer (2013) tested participants on a same/different task similar to that of Pilz et al. (2009) but using black and white images. They found that the shift from view-dependent to view-independent face recognition occurred suddenly. In other words, faces appeared to be categorised as either familiar or unfamiliar. However, their pairs of view-dependent images were identical, so they probably tested image rather than face recognition. Zimmermann and Eimer also noticed this change after a longer experimental break, so learning could have consolidated during the break while it was not being measured. These studies show the difficulty with generalising and consolidating across studies. So, the present study aims to conduct a broad investigation into experimental face learning using a variety of face types, and a variety of behavioural and physiological measures.

While most researchers have examined face recognition using explicit decision processes (e.g. familiar/unfamiliar judgements or matching tasks) researchers have found various other ways to test it, such as using neurological responses (e.g. Rossion et al., 2001; Caldara & Abdi, 2006; or see Gobbini & Haxby, 2007, for a review). One area that has proved fruitful is eye-tracking, which can involve the detection and recording of fixations and saccades as participants look at faces. Such measures can reveal which parts of the face appear to be more important for recognition (Van Belle, 2010; Hills & Pake, 2013) and track gaze as people view faces (Barton, Radcliffe, Cherkasova, Edelman, & Intriligator, 2006). Eye-tracking equipment can also record pupillary changes and blinks, which have been used to make inferences about cognitive load (Chen & Epps, 2014), primarily in reading (e.g. Schluroff et al., 1986) or mathematics tasks (Jainta & Baccino, 2010). Cognitive load describes the amount of mental effort required to do a task, and pupillary analysis has shown that pupil sizes are larger when people are doing difficult

tasks, and are thus associated with greater cognitive load (Piquado, Isaacowitz, & Wingfield, 2010).

The present study aims to investigate whether physiological responses provide an indirect index of face learning that is more informative than decision responses or reaction times (RTs), by recording these two behavioural measures, as well as fixations, blinks and pupillary responses. The focus will be on pupillary responses, as they are associated with cognitive load (Chen & Epps, 2014). We aim to investigate whether pupillary responses change as faces are learnt, and whether this co-occurs with changes in accuracy rates. First, we predict that familiar/unfamiliar decision responses will become more accurate as faces are learnt. Second, since familiar face recognition is easier than unfamiliar face recognition, familiar faces should require less effort to process. Therefore, we predict that pupils will be larger when processing unfamiliar faces than familiar faces, and that pupils will get smaller as faces are learnt (Goldinger, He, & Papesh, 2009). Pupillary responses should mirror improvements in decision responses (accuracy) if they are reliable indices of overt face learning. Also, they could potentially index implicit face learning better than explicit decision responses, as the latter can be contaminated by conscious decision-making processes required to make the responses (such as motivation to make one response over another, or an error such as the wrong key-press). We will investigate RTs, fixations and blinks in the same way, and anticipate that they too will diminish during the experiment. It is also predicted that the process of face learning will be gradual rather than categorical, in line with previous studies (Tong & Nakayama, 1999; Kosaka et al., 2003; and Pilz et al., 2009). In other words, it is expected that faces lie on a spectrum of familiarity rather than being categorised as either familiar or unfamiliar.

Finally, we wanted to investigate the effects of age on face learning; whether face learning declines with age, or whether own-age faces were learnt more quickly, successfully, or easily than other-age faces. We predict that own-age faces would be learnt and classified more easily than other-age faces.

Experiment 1

2.2. Method

2.2.1. Design

This study used a mixed design: repeated measures on *trial block* (with six trial blocks: 1, 2, 3, 4, 5 or 6) *familiarity* (with two familiarity types: *familiar* and *unfamiliar*), *face age* (with two face types: *young faces*, and *old faces*), *face gender* (with two face types: *male faces* and *female faces*), and independent measures on *participant gender* (with two genders: *male* and *female*) and *participant age* (with two age groups: *young participants* and *old participants*). The dependent variables were decision responses (accuracy), reaction times (RTs), number of fixations, number of blinks and pupil sizes (see section 2.3.3.1.).

2.2.2. Participants

Thirty-nine participants with normal or corrected to normal vision were recruited either via the university, or from the local community. All participants participated in both experimental sets. Seventeen (six males and eleven females) were Caucasian local community members aged between 68 and 75, and twenty-two (eleven male and eleven female) were Caucasian university students aged between 18 and 27. We also recruited five further elderly males, but these had to be excluded due to either technical issues,

failure to calibrate either eye, or because they misunderstood the task. We had hoped to recruit forty participants (ten from each gender/age group), but had extreme difficulty recruiting older participants. By using participants from a pilot experiment, we had usable data from eleven older females, as well as those from three young females and three young males from the pilot experiment, but were unable to recruit enough older males. This resulted in the number of participants described above.

2.2.2. Apparatus and Materials

There were two experimental sets containing different face types (see fig. 5 for diagram of procedure and examples of stimuli). The *young* set consisted of young Caucasian faces aged approximately 18-30 years. The *old* set consisted of Caucasian faces aged approximately 60-75 years. Each face set contained equal numbers of male and female faces. For each face set, there were familiarisation stimuli. These consisted of three colour five-second silent video clips of talking faces. The video stimuli were taken from VidTIMIT (2009) in JPEG image format (512 x 384 pixels) and converted to XVID (using SplitAvi) for compatibility with Experiment Builder. Video clips were counterbalanced: the three video stimuli were shown in a different order for each participant, as were the face sets: some participants saw the male faces first, while some saw the female faces first. The order of the images in the test phase was randomised. The stimuli were displayed at an approximate distance of 60cm from the chin rest, as this is the optimum distance for recording pupillary responses.

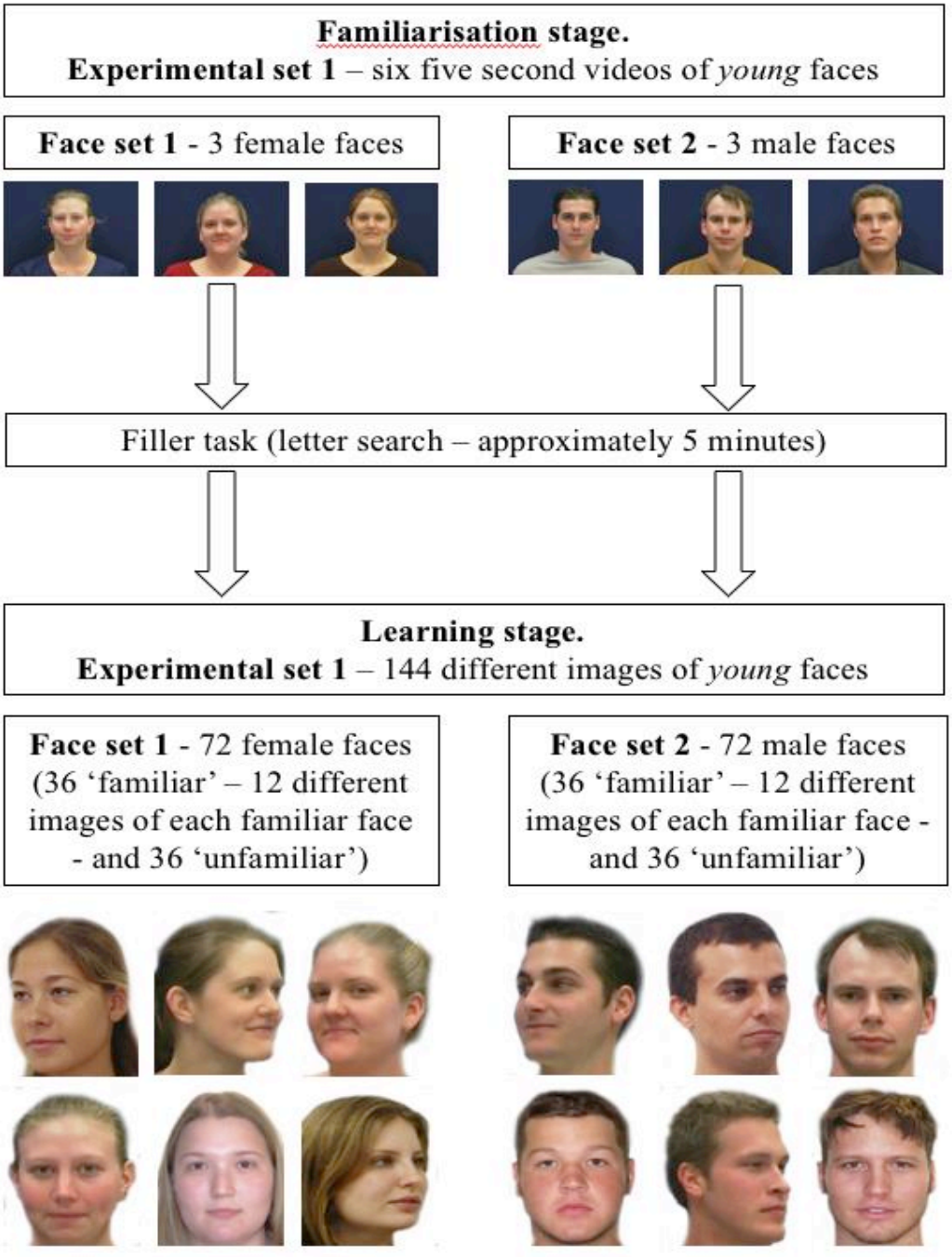
The learning stage included 36 different images of the previously-seen individuals (*familiar*) that were matched to 36 images of novel faces (*unfamiliar*). The *familiar* stimuli were taken from VidTIMIT (2009). The unfamiliar images were taken from VidTIMIT and the FEI Face Database (2006). All images were unique and showed twelve

frontal, twelve three-quarter left and twelve three-quarter right viewpoints. The images were cropped and matched for size (20 X 25 cm), resolution (161 X 229 pixels) and luminance. They consisted of full colour headshots against a white background. Distinguishing items such as moles, scars and piercings were removed. Image luminance was verified using a photometer app (myLightMeter, 2015), that was held against the screen to produce a measurement of incident light metering (light falling on the screen) and reflective light metering (light reflected off the screen). This was done to control for pupillary fluctuations related to changing luminance.

Experiment Builder was run on a 21.5 inch iMac computer and a video desktop EyeLink 1000 eye tracker, which uses an infrared camera. The head was stabilised using a chin rest, although the EyeLink can accommodate small head movements, wobble and blinks. The right eye was tracked for all participants.

The gaze and pupillary recordings were calculated in the following ways. The camera shines an invisible infrared light into the eye that hits the back of the cornea, causing a reflection. The distance between this reflection and the centre of the pupil is used to measure changes in gaze. Pupil size is determined by covering the pupil with blue pixels, which are counted by the software to produce a pupil size score. This is usually converted into a meaningful score such as a percentage, and this is the approach that we used.

Example of stimuli presentation for Experimental set 1, young faces. The same procedure was used for Experimental set 2, old faces.



Note: The Face sets were shown separately and counter-balanced

Fig. 5. Example of stimuli and procedure for experiment 1

2.2.4. Procedure

The participants were briefed as to the aims and procedure of the experiment. They then placed their chins in the chin rest and their eye movements were calibrated to nine points on the display. At this point three video clips were played in the familiarisation stage. These were followed by further instructions, then a drift check that required the participant to look at a black dot on a white screen. The drift check checks that gaze accuracy is maintained throughout the experiment. This was followed by a short filler task (a word search), then by the learning stage, containing 72 faces. Of these, 36 were 'familiar' and contained different images of one of the three individuals seen in the previous video clips, and 36 were 'unfamiliar', novel faces that had been matched to the faces in the video clips, (based on physical appearance such as skin tone, hair style etc.). The 72 faces were displayed sequentially in a random order, with a drift check between successive images. The participant was asked to click 'F' if the face looked familiar and 'U' if it appeared unfamiliar. Each image was displayed until the participant had responded. The ISI (interstimulus interval) was approximately 2-3 seconds, the amount of time for the participant's eye to stabilise on the black dot of the drift check, and for the experimenter to press the space bar that triggered presentation of the next image. This was also sufficient time for pupil sizes to 're-set' between images. The eye-tracker recorded eye movements and pupillary responses as the participant viewed the video clips and images. This procedure was repeated four times for each participant, once for each face type (young male, young female, old male and old female).

2.3. Results

As there were many interactions in this experiment, we created two graphs for each DV that expressed the data we were most interested in investigating: trial block, participant age and gender, and familiarity. Therefore, each section has one graph presenting the data for familiar faces and one presenting the data for unfamiliar faces over the six sequential trial blocks, both as a function of participant age and gender. Any further interactions are presented in additional graphs.

2.3.1. Decision Responses (Accuracy)

The decision response for each image was recorded, and the mean percentage of correct scores was calculated for each of the six trial blocks by following the procedure below.

2.3.1.1. Trial blocks

In each face set there were three individuals, and in the learning phase there were 12 different image trials for each of these faces, totalling 36 familiar face trials. In order to track any changes in accuracy rates as the experiment progressed, these 36 images were grouped into six successive trial blocks, each containing six images: the first six faces seen were allocated to the first trial block, the second six faces were allocated to the second trial block, and so on. We obtained one score from each of these trial blocks, by calculating the mean accuracy score from each of the six images allocated to it. The same procedure was conducted for the unfamiliar faces. We repeated this procedure for each dependent variable.

A Mixed ANOVA was performed to compare accuracy while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 33.06, p = .01. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .67).

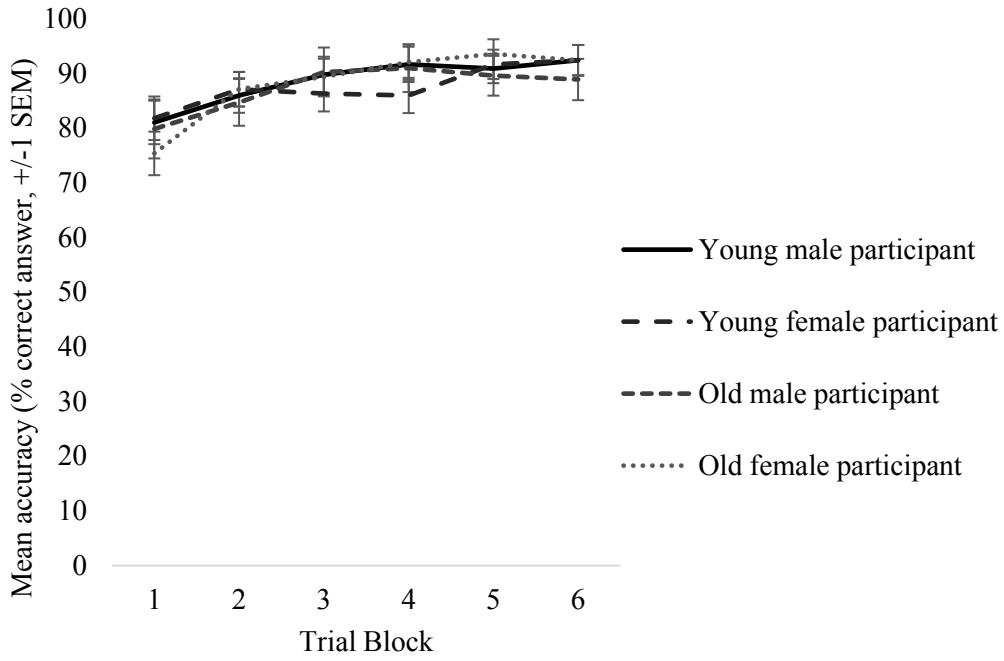


Fig. 6. Mean accuracy for *familiar* faces over six sequential trial blocks, as a function of participant age and gender.

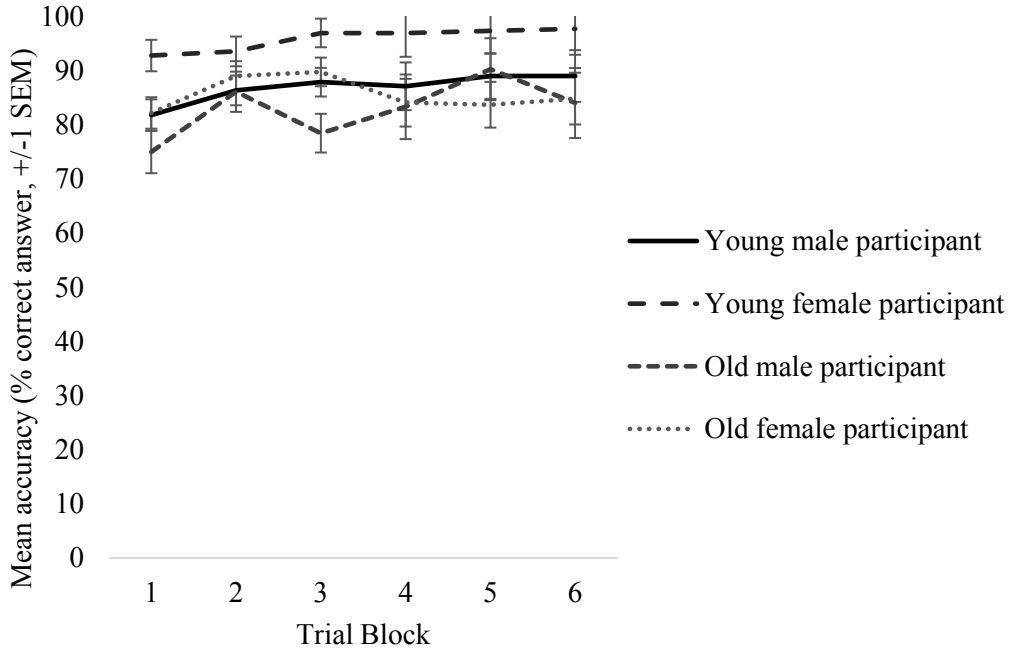


Fig. 7. Mean accuracy for *unfamiliar* faces over six sequential trial blocks, as a function of participant age and gender.

As shown in figs. 6 and 7, there was a significant effect of *trial block*, F (3.35, 117.20) = 17.85, $p < .001$, $r = .36$, $\eta^2 = .34$. Planned contrasts revealed that participants were significantly more accurate in the second trial block compared to the first overall ($p < .001$), but other comparisons were not significant. There was a no effect of *participant age, F* (1, 35) = 3.67, $p = .06$ (young: $M = 89.98$, $SE = 1.29$; old: $M = 86.47$, $SE = 1.47$).

As can be seen in fig. 8, there was also a significant interaction between *familiarity* and *face age*, F (1, 35) = 4.44, $p = .04$, $\eta^2 = .12$ (familiar young face: $M = 88.47$, $SE = 1.44$; familiar old face: $M = 87.46$, $SE = 1.94$; unfamiliar young face: $M = 85.54$, $SE = 1.87$; unfamiliar old face: $M = 90.07$, $SE = 2.01$).

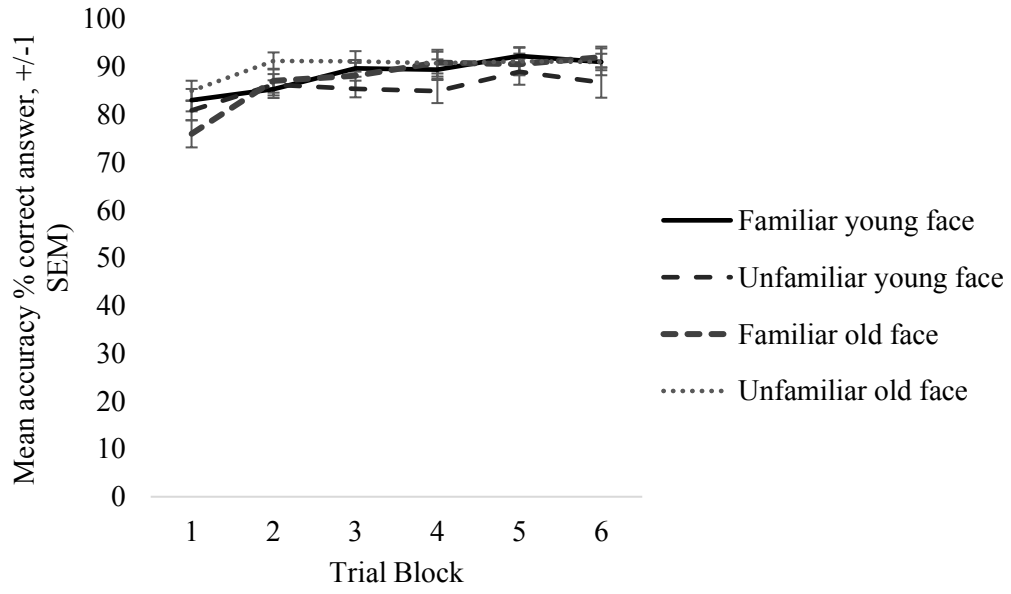


Fig. 8. Mean accuracy for all faces over six sequential trial blocks, as a function of face age and familiarity.

Analysis of the results suggested that while participants made decisions about familiar faces of both age groups similarly accurately, they differed in accuracy when making decisions about unfamiliar faces of different ages. They appeared to be least

accurate classifying unfamiliar young faces as unfamiliar and most accurate classifying unfamiliar old faces as unfamiliar.

2.3.2. Reaction Times (ms)

The reaction time (RT) for each image was recorded, and the mean RT was calculated for each trial block.

A similar analysis was performed to analyse the reaction time data. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 121.93, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .43$).

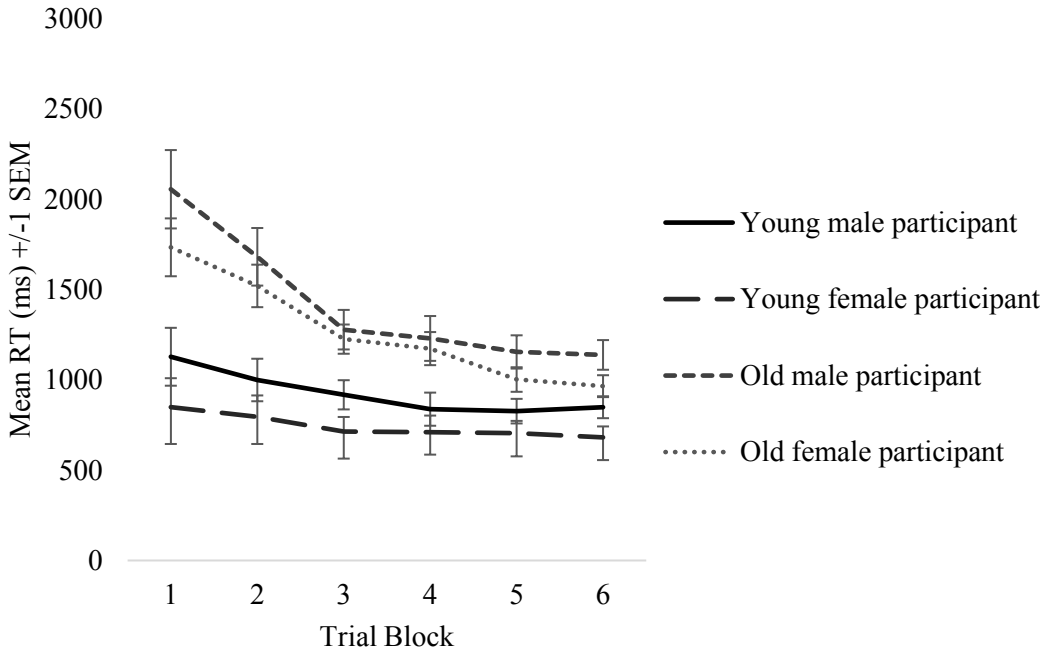


Fig. 9. Mean RT (ms) for *familiar* faces over six sequential trial blocks, as a function of participant age and gender.

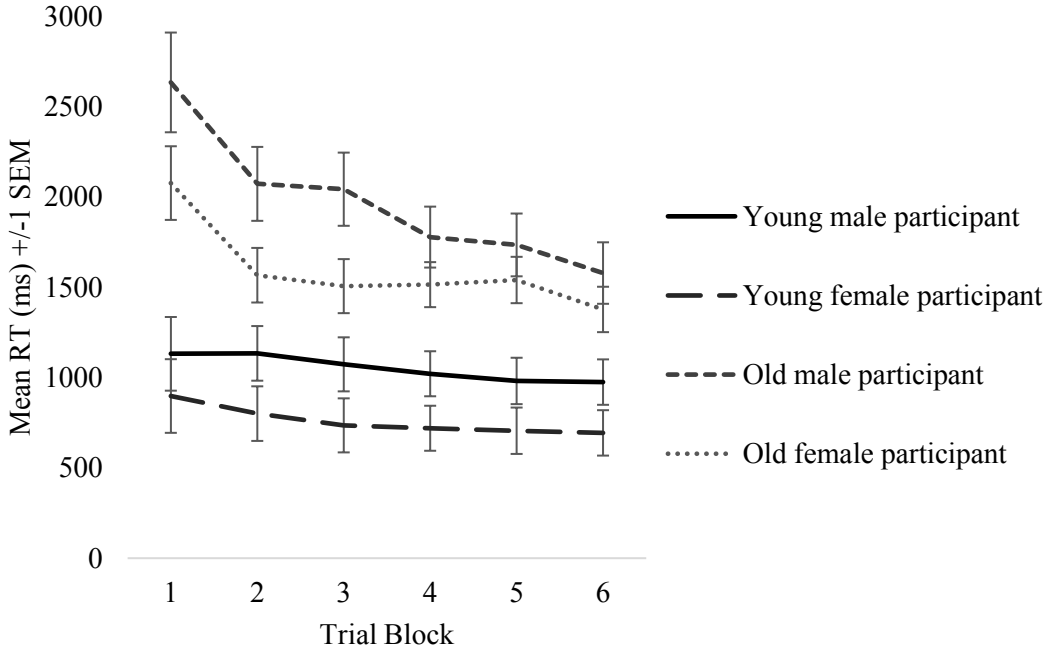


Fig. 10. Mean RT (ms) for *unfamiliar* faces over six sequential trial blocks, as a function of participant age and gender.

As shown in figs. 9 and 10, there were significant effects of *trial block*, F (2.15, 75.23) = 51.19, $p < .001$, $r = .64$, $\eta^2 = .59$. Planned contrasts revealed that in young participants RTs became significantly shorter between the first and second trial blocks ($p = .02$), the second and the third ($p < .001$) and the third and fourth ($p <. 05$). In old participants RTs became significantly shorter between the first and second trial blocks ($p = .01$), the second and the third ($p = .01$) and the fifth and sixth ($p =. 03$). No other comparisons were significant.

There were also significant effects of *familiarity*, F (1, 35) = 21.95, $p < .001$, $r = .62$, $\eta^2 = .39$ (familiar face: $M = 1090.69$, $SE = 46.76$; unfamiliar face: $M = 1345.73$, $SE = 74.45$); *participant age*, F (1, 35) = 38.75, $p < .001$ (young participant: $M = 870.29$, $SE = 71.91$; old participant: $M = 1566.12$, $SE = 85.59$); and *participant gender* F (1, 35) = 5.07, $p = .03$ (male participant: $M = 1344.08$, $SE = 85.59$; female participant: $M = 1092.33$, $SE = 71.91$).

Figs. 9 and 10 also show the significant interactions between *trial block* and *participant age*, F (2.15, 75.23) = 18.15, $p < .001$, $\eta^2 = .34$, and between *familiarity* and *participant age*, F (1, 35) = 11.41, $p < .001$, $\eta^2 = .25$ (familiar face, young participant: $M = 834.69$, $SE = 60.16$; unfamiliar face, young participant: $M = 905.90$, $SE = 95.78$: familiar face, old participant: $M = 1346.70$, $SE = 71.60$; unfamiliar face, old participant: $M = 1785.55$, $SE = 114.00$).

Results also showed an effect of *face age*, F (1, 35) = 15.88, $p < .001$, $\eta^2 = .31$ (young face: $M = 1292.08$, $SE = 65.16$; old face: $M = 1144.33$, $SE = 51.86$), and as shown in fig. 11, there was an interaction between *face age* and *participant age*, F (1, 35) = 15.88, $p < .001$, $\eta^2 = .16$ (young face, young participant: $M = 896.10$, $SE = 83.83$; old

face, young participant: M = 844.49, SE = 66.78: young face, old participant: M = 1688.07, SE = 99.78; old face, old participant: M = 1444.18, SE = 79.41).

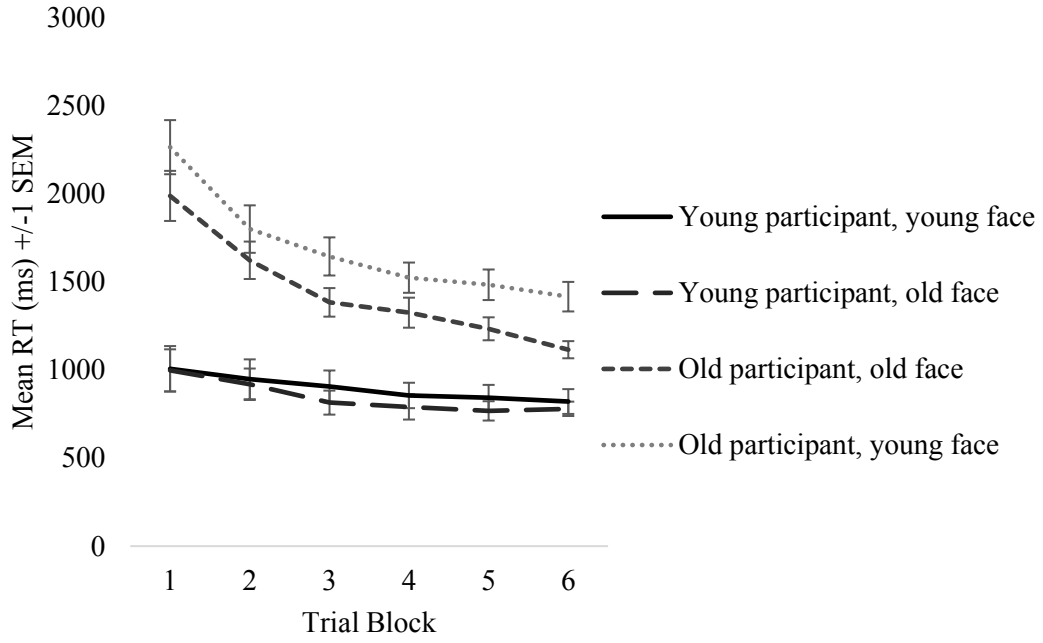


Fig. 11. Mean RT (ms) for all faces over six sequential trial blocks, as a function of participant age and face age.

As shown in Table 1, there was also a three-way interaction between *familiarity*, *face age* and *face gender*, F (1, 35) = 7.66, p = .01, η^2 = .18.

Table 1. Three-way interaction, showing data for young male, young female, old male and old female faces as a function of familiarity (familiar and unfamiliar)

	Young male face	Young female face	Old male face	Old female face
Familiar	*M* = 1171.794 (*SE* = 59.30)	*M* = 1112.111 (*SE* = 62.72)	*M* = 1004.979 (*SE* = 37.93)	*M* = 1073.885 (*SE* = 62.47)
Unfamiliar	*M* = 1370.692 (*SE* = 79.22)	*M* = 1513.74 (*SE* = 104.34)	*M* = 1244.144 (*SE* = 84.20)	*M* = 1254.324 (*SE* = 63.77)

There were also some four and five-way interactions. These included: *participant age*, *familiarity*, *face age* and *face gender*, F (1, 35) = 5.66, p = .02, η^2 = .14; *participant age*, *participant gender, face age, face gender and trial block*, F (5, 175) = 5.66, p = .02; *participant gender, familiarity*, *face age, face gender and trial block*, F (3, 175) = 4.31, p = .01, η^2 = .10; and one six-way interaction: *participant age*, *participant gender, familiarity*, *face age, face gender and trial block*, F (5, 175) = 4.16, p = .02, η^2 = .11.

The results indicated that people looked longer at unfamiliar faces than familiar faces, and that RTs decreased with each successive trial block. This occurred more dramatically in old participants than young participants. The results also showed that old participants reacted more slowly than young participants, and that participants reacted more slowly to young faces than old faces. Also, when old people looked at old faces, they reacted more quickly than when looking at young faces (the opposite effect was absent in young participants). Finally, participants looked longest at unfamiliar young female faces.

2.3.3. Pupillary responses

2.3.3.1. Pupil sizes

The Eyelink 1000 (n.d.) produces an arbitrary figure to represent pupil size, based on the number of pixels. Therefore, it is important to convert this figure into something meaningful. There is no agreed consensus on an appropriate method, although some researchers us a pupil dilation ratio (e.g. Otero, Weekes & Hutton, 2011). The method chosen here was to use percentages, as these are calculated in a similar way, are readily understood, and easy to interpret. To do this, we started with the mean pupil size for each image that was provided by the eye tracker. For each participant, percentages were calculated separately by identifying the image with the largest mean pupil size (calculated as 100%), and the image with the smallest mean pupil size (calculated as 0%). The mean pupil size for each other image was then calculated as a percentage of the difference between the two. The mean pupil size percentage was then calculated for each trial block.

A similar analysis was performed to analyse pupil sizes while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 147.52, $p > .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .31$).

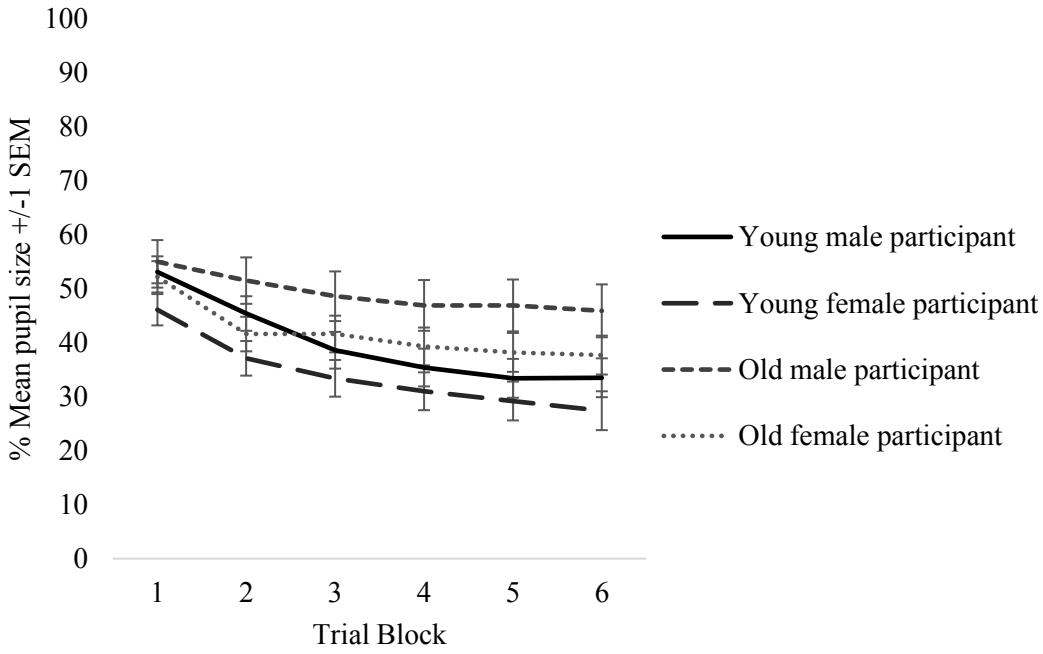


Fig. 12. Mean pupil sizes for *familiar* faces over six sequential trial blocks, as a function of participant age and gender.

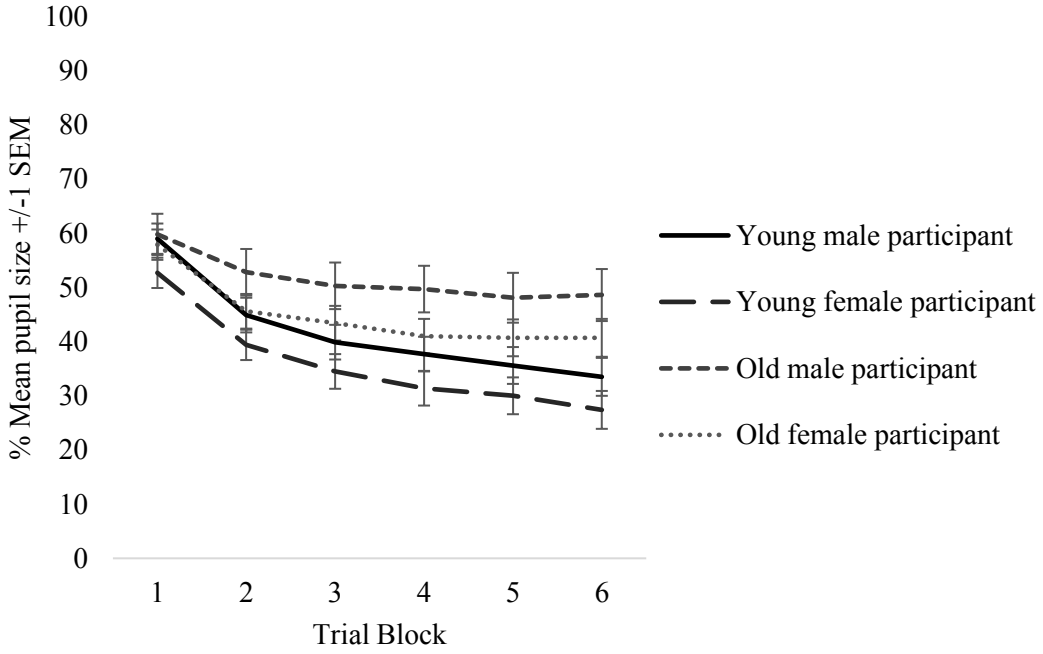


Fig. 13. Mean pupil sizes for *unfamiliar* faces over six sequential trial blocks, as a function of participant age and gender.

As shown in figs. 12 and 13, the results show that there were significant effects of *trial block*, F (1.55, 54.14) = 106.24, $p < .001$, $r = .81$, $\eta^2 = .75$; where planned contrasts revealed that in young participants each trial block elicited a significantly smaller pupil size than the preceding trial block (all ps < .001), between the fourth and fifth ($p = .01$) apart from between the fifth and sixth ($p = .07$), while in old participants only the first two comparisons elicited significantly smaller pupil sizes ($p < .001$, and $p = .01$ respectively).

There were also significant effects of *familiarity*, F (1, 35) = 28.21, $p < .001$, r = .67, $\eta^2 = .44$ (familiar face: $M = 41.30$, $SE = 1.75$; unfamiliar face: $M = 43.54$, $SE = 1.63$), *participant age*, F (1, 35) = 7.27, p =.01, r = .41 (young participant: M = 37.90, SE = 2.16; old participant: $M = 46.94$, $SE = 2.57$).

Figs. 12 and 13 also show that there were interactions between *trial block* and *familiarity, F* (4.34, 151.88) = 10.10, $p < .001$, $\eta^2 = .23$, and between *trial block* and *participant age*, F (1.55, 54.14) = 7.08, $p < .001$, $\eta^2 = .17$.

As seen in fig. 14, there was also an interaction between *face gender* and *face age*, F (1, 35) = 6.64, $p = .01$, $\eta^2 = .16$ (young male face: $M = 42.5$, $SE = 2.30$; old male face: M = 42.90, SE = 2.00; young female face: M = 39.90, SE = 2.10; old female face: M = 44.40, $SE = 2.10$), although the differences were very small. No other interactions reached significance.

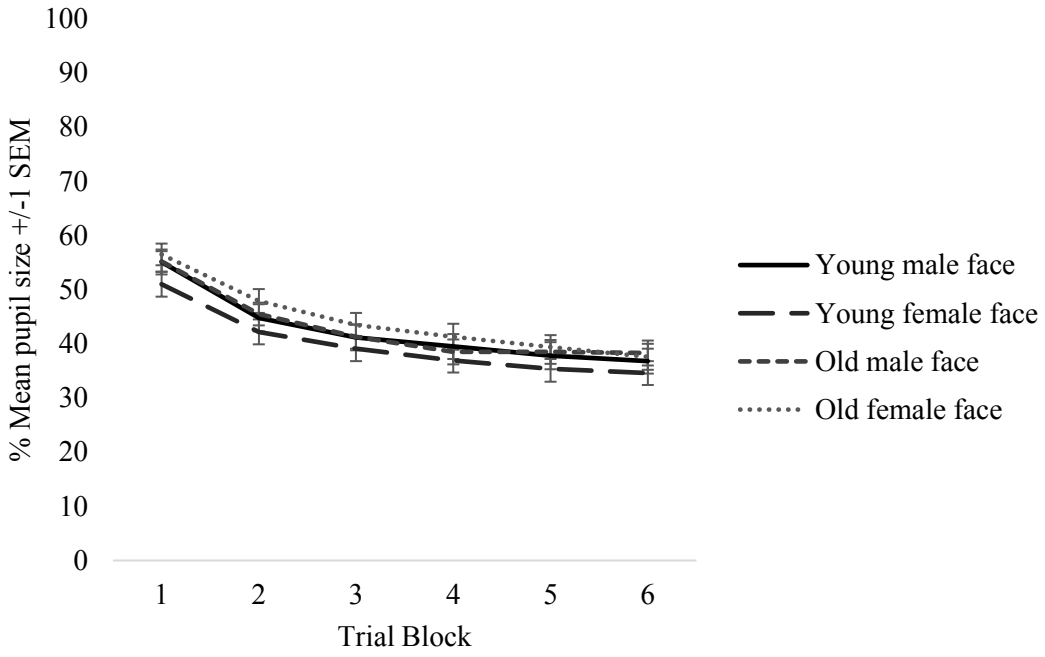


Fig. 14. Mean pupil sizes in all participants over six sequential trial blocks, as a function of face age and gender.

Analysis of the results indicated that participants had larger pupil sizes when looking at unfamiliar faces, that pupil sizes decreased in each successive trial block, with a steeper initial decrease at the start of the experiment when looking at unfamiliar faces. Old participants had significantly larger pupil sizes than young participants. Old participants' pupil sizes also decreased less during the experiment. Finally, overall pupils were similar sizes when looking at male faces, smallest when looking at young female faces, and largest when looking at old female faces.

Fixations

The mean number of fixations for each trial was calculated, and the mean fixation score was calculated for each trial block.

A similar analysis was performed to analyse the number of fixations while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 116.08, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .39$).

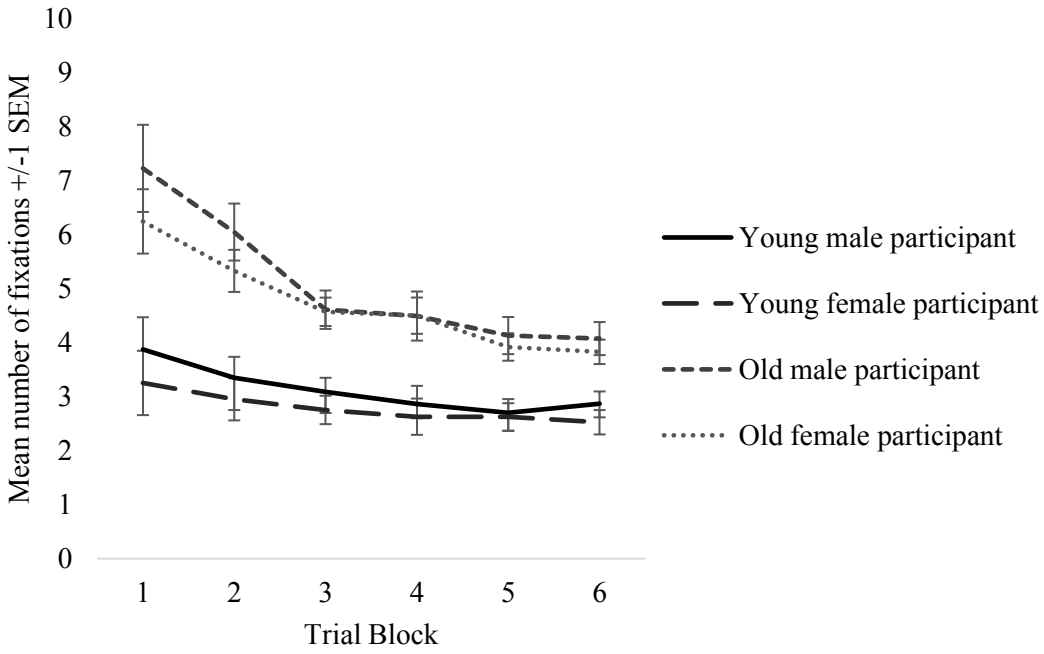


Fig. 15. Mean number of fixations for *familiar* faces over six sequential trial blocks, as a function of participant age and gender.

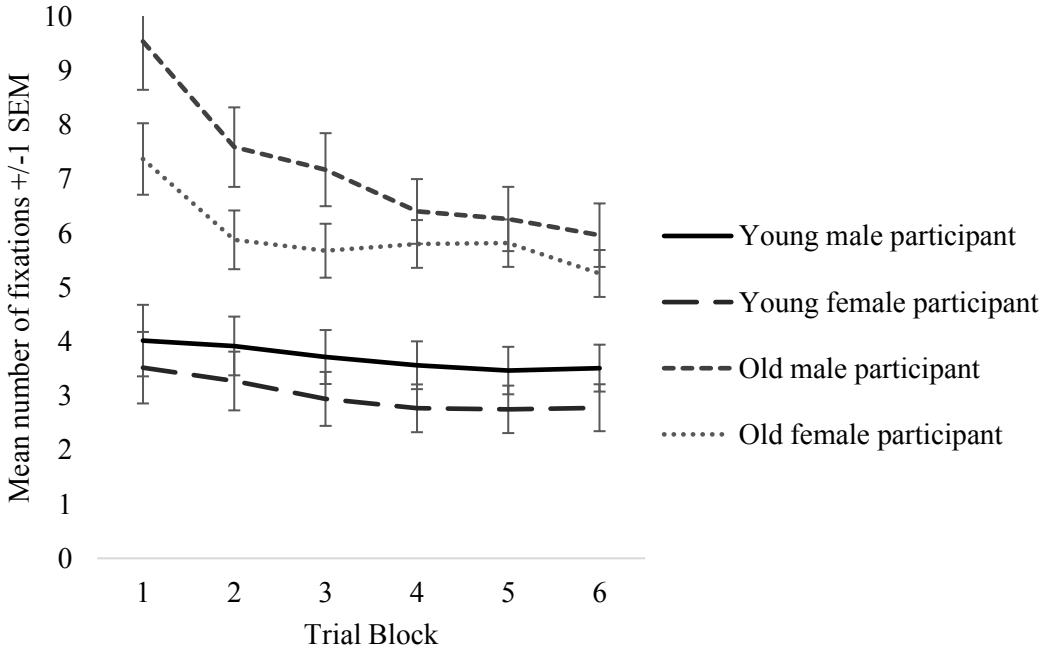


Fig. 16. Mean number of fixations for *unfamiliar* faces over six sequential trial blocks, as a function of participant age and gender.

As seen in figs. 15 and 16, there were significant effects of *trial block*, F (1.95, 68.30) = 52.16, $p < .001$, $r = .66$, $\eta^2 = .60$, where planned contrasts revealed that young participants made significantly fewer fixations between the first and second trial block ($p < .001$), the second and third ($p = .01$), and between the third and fourth ($p = .02$). Old participants also made significantly fewer fixations between the first and second ($p < .001$) and the second and third ($p = .01$). No other comparisons were significant.

There was also a significant effect of *familiarity*, F (1, 35) = 29.09, $p < .001$, $r = .67$, $\eta^2 = .45$, (familiar face: $M = 3.94$, $SE = 0.17$; unfamiliar face: $M = 4.96$, $SE = 0.26$), *face age*, F (1, 35) = 23.33, $p < .001$ (young face: $M = 4.77$, $SE = 0.23$; old face: $M = 4.13$, $SE = 0.19$), and *participant age*, F (1, 35) = 42.31, $p < .001$ (young participant: $M = 3.16$, $SE = 0.26$; old participant: $M = 5.74$, $SE = 0.30$).

These figs also show that were interactions between *trial block* and *participant age*, F (1.95, 68.30) = 15.71, $p < .001$, $\eta^2 = .31$, *familiarity* and *participant age*, F (1, 35) = 10.92, $p = .01$ (familiar face, young participant: $M = 2.96$, $SE = 0.22$; unfamiliar face, young participant: $M = 3.35$, $SE = 0.33$; familiar face, old participant: $M = 4.91$, $SE = 0.26$; unfamiliar face, old participant: $M = 6.56$, $SE = 0.40$),

As seen in fig. 17, interactions were also found between *familiarity* and *face age*, F (1, 35) = 4.48, $p = .04$, $\eta^2 = .11$ (familiar young face: $M = 4.17$, $SE = 0.21$; unfamiliar young face: $M = 5.36$, $SE = 0.29$; familiar old face: $M = 3.70$, $SE = 0.16$; unfamiliar old face: $M = 4.56$, $SE = 0.25$).

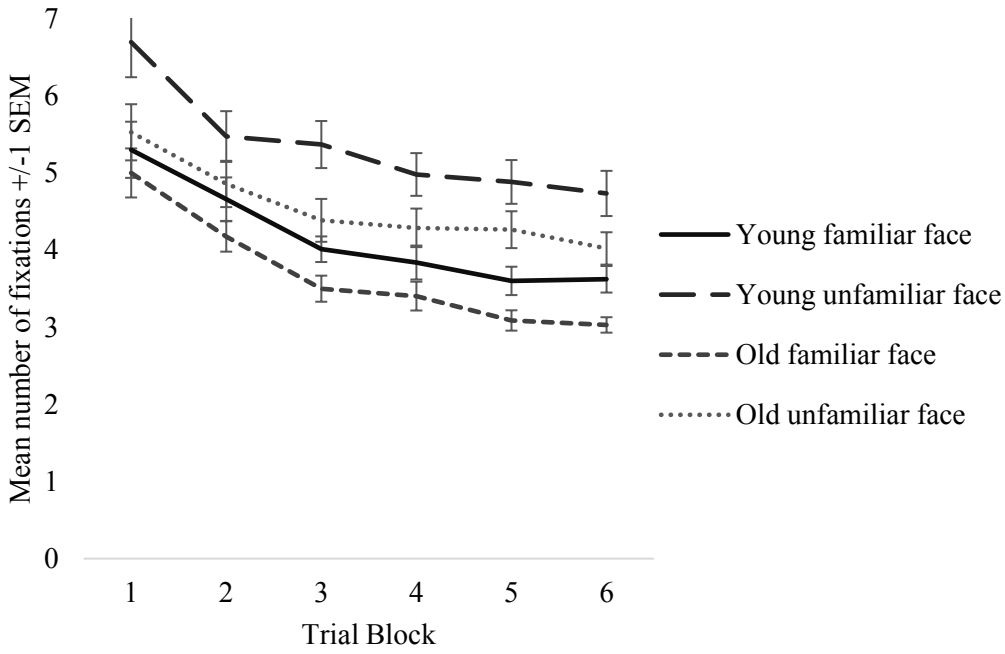


Fig. 17. Mean number of fixations for all faces over six sequential trial blocks, as a function of familiarity and face age.

As seen in fig 18, there was also an interaction of *face age* and *participant gender*, F (1, 35) = 5.64, p = .02, η^2 = .14 (young face, male participant: M = 5.24, SE = 0.35; young face, female participant: M = 4.29, SE = 0.30; old face, male participant: M = 4.30, SE = 0.29; old face, female participant: M = 3.96, SE = 0.24): the highest number of fixations were made when males looked at young faces.

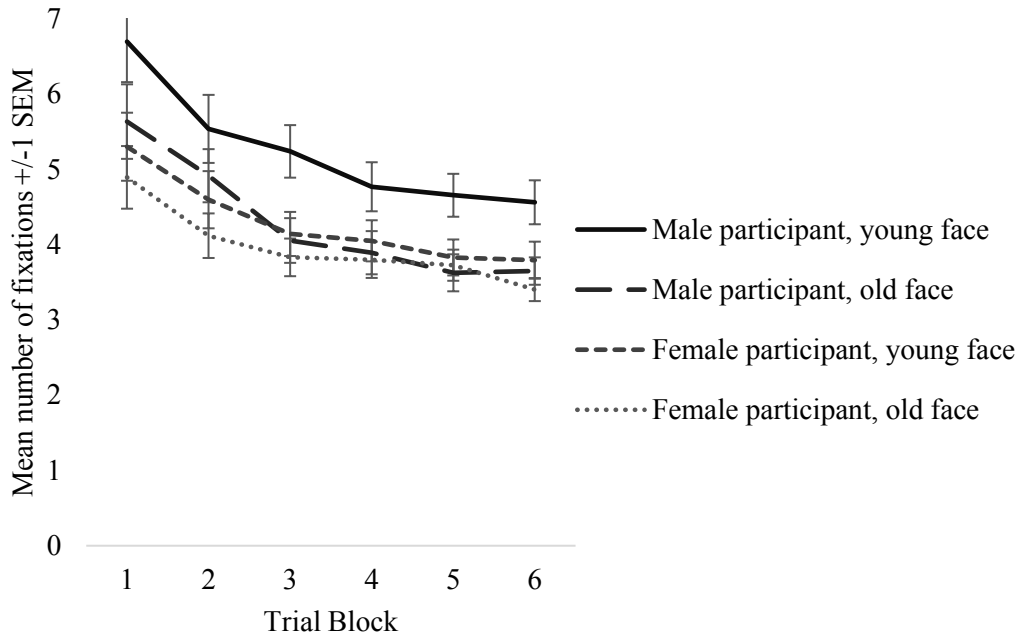


Fig. 18. Mean number of fixations for all faces over six sequential trial blocks, as a function of participant gender and face age.

As seen in fig. 19. there was also an interaction between *face age* and *participant age*, F (1, 35) = 8.27, p = .01, η^2 = .19 (young face, young participant: M = 3.28, SE = 0.30; young face, old participant: M = 6.25, SE = 0.35; old face, young participant: M = 3.03, SE = 0.24; old face, old participant: M = 5.23, SE = 0.29).

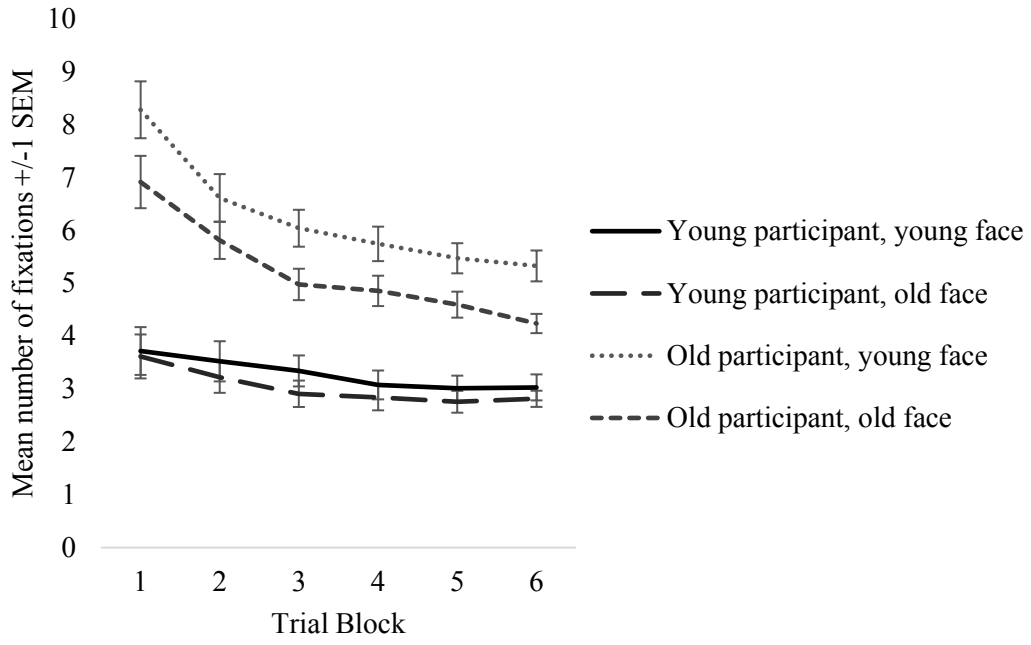


Fig. 19. Mean number of fixations for all faces over six sequential trial blocks, as a function of participant age and face age.

Also, as seen in Tables 2 and 3, there were three-way interactions between *familiarity*, *face age* and *face gender*, F (1, 35) = 4.01, p = .05, η^2 = .10, and between *familiarity*, *face age* and *participant gender*, F (1, 35) = 4.14, p = .05, η^2 = .11.

Table 2. Three-way interaction, showing data for young male, young female, old male and old female faces, as a function of familiarity (familiar and unfamiliar)

	Young male face	**Young female face**	**Old male face**	**Old female face**
Familiar face	*M* = 4.28 (*SE* = 0.22)	*M* = 4.07 (*SE* = 0.23)	*M* = 3.58 (*SE* = 0.15)	*M* = 3.82 (*SE* = 0.21)
Unfamiliar face	*M* = 5.22 (*SE* = 0.30)	*M* = 5.49 (*SE* = 0.33)	*M* = 4.48 (*SE* = 0.30)	*M* = 4.64 (*SE* = 0.22)

Table 3. Three-way interaction, showing data for familiar young, unfamiliar young, familiar old and unfamiliar old faces, split by participant gender (male and female)

	Familiar young face	**Unfamiliar young face**	**Familiar old face**	**Unfamiliar old face**
Male participant	*M* = 4.43 (*SE* = 0.32)	*M* = 6.06 (*SE* = 0.45)	*M* = 3.80 (*SE* = 0.24)	*M* = 4.79 (*SE* = 0.39)
Female participant	*M* = 3.92 (*SE* = 0.27)	*M* = 4.65 (*SE* = 0.38)	*M* = 3.60 (*SE* = 0.20)	*M* = 4.32 (*SE* = 0.32)

There were also some four and five-way interactions. These included: *participant age*, *participant gender, familiarity* and *face age*, F (1, 35) = 4.85, p = .03, η^2 = .12; *participant age, face age, face gender and familiarity*, F (1, 35) = 4.91, p = .03, η^2 = .13; *participant age*, *familiarity*, *face age, and trial block*, F (5, 175) = 3.44, p = .01, η^2 = .09; *participant gender, familiarity*, *face age and trial block*, F (5, 175) = 3.40, p = .01, η^2 = .09; and *participant age*, *participant gender, face gender and trial block*, F (5, 175) = 2.73, p = .02, η^2 = .03.

Unfamiliar faces elicited significantly more fixations than familiar faces, and the number of fixations gradually decreased with each successive trial block. Old participants made more fixations than young participants, and across trial blocks, the number of fixations decreased more dramatically in old participants than young participants. When old participants looked at unfamiliar faces, they made almost twice as many fixations as

young participants did (and this difference was greater than when both age groups looked at familiar faces). Similarly, when young participants looked at the faces of both age groups they made similar numbers of fixations, regardless of face age, but old participants made considerably more fixations to other-age faces than they did to own-age faces. Also, when faces were young, female and unfamiliar they elicited the highest number of fixations, but when faces were familiar the young male faces did. Finally, the highest number of fixations were made when males looked at unfamiliar young faces of both genders.

2.3.4. Blinks

For each participant, the mean number of blinks for each trial was calculated, and then the mean blink score was calculated separately for each trial block.

A similar analysis was performed to analyse the number of blinks while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 92.47, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .51$).

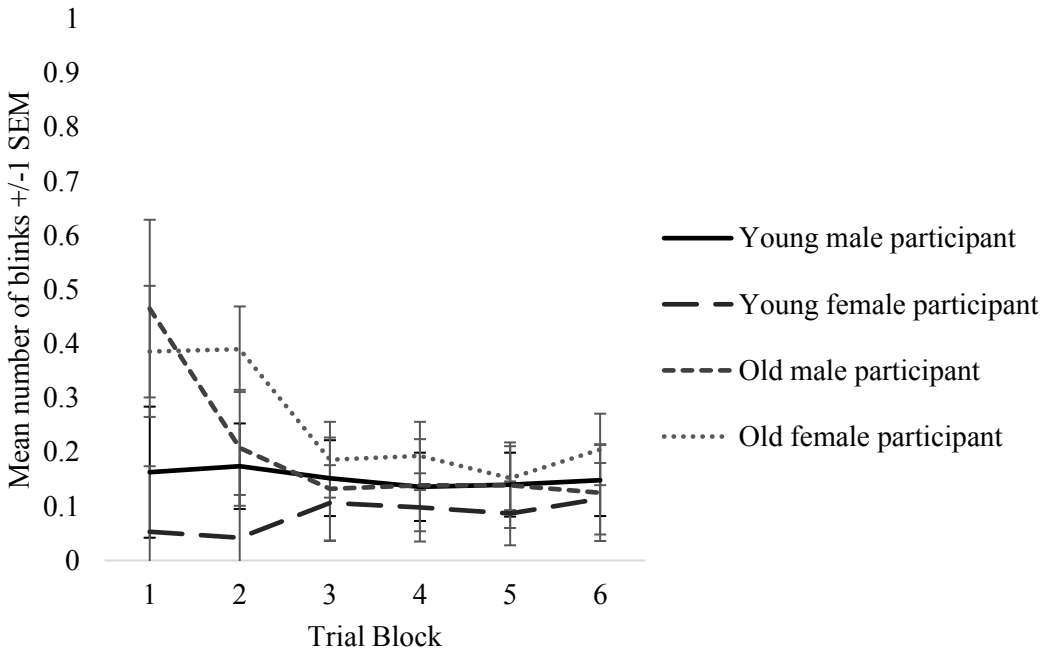


Fig. 20. Mean number of blinks for *familiar* faces over six sequential trial blocks, as a function of participant age and gender.

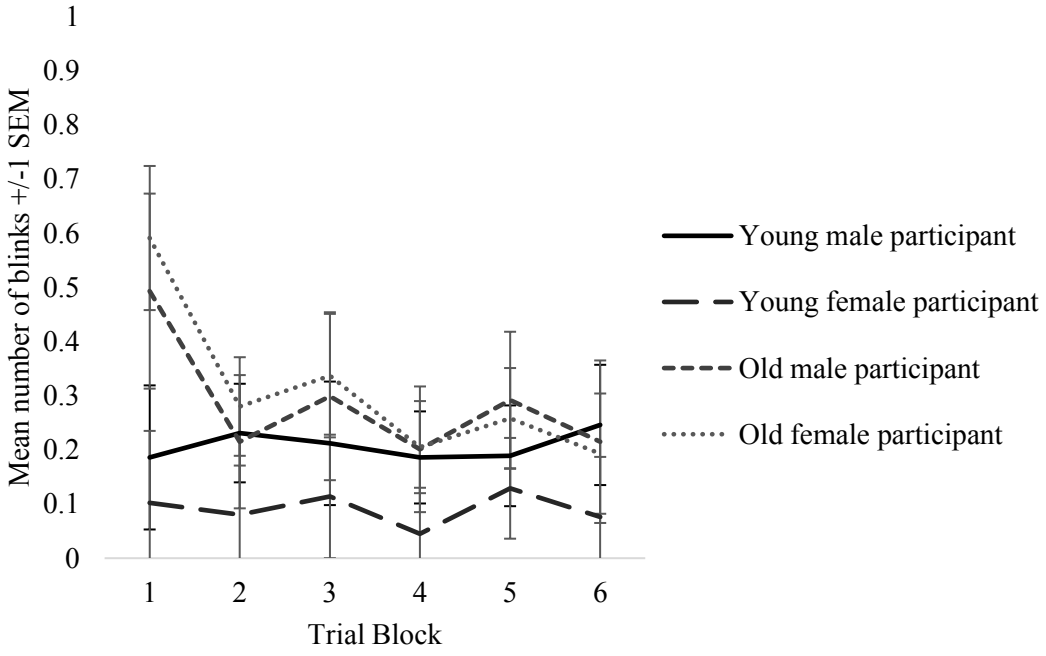


Fig. 21. Mean number of blinks for *unfamiliar* faces over six sequential trial blocks, as a function of participant age and gender.

As seen in figs. 20 and 21, the results show that there were significant effects of *trial block*, F (2.55, 89.13) = 6.74, $p < .001$, $r = .27$, $\eta^2 = .16$; where planned comparisons revealed that the second trial block had significantly fewer blinks than the first ($p = .02$) and that the fourth elicited significantly fewer blinks than the third ($p < .05$), *familiarity*, F (1, 35) = 4.94, $p = .03$, $r = .35$ (familiar face: $M = 0.17$, $SE = 0.3$; unfamiliar face: $M = 0.23$, $SE = 0.5$), and an interaction between *trial block* and *participant age*, F (2.55, 89.12) = 7.51, $p < .001$, $\eta^2 = .18$.

As shown in Table 4, results also found a three-way interaction between *familiarity*, *face age* and *face gender*, F (1, 35) = 7.23, $p = .01$, $\eta^2 = .18$, but no other significant interactions.

Table 4. Three-way interactions, showing data for young male, young female, old male and old female faces, as a function of familiarity (familiar and unfamiliar)

	Young male face	**Young female face**	**Old male face**	**Old female face**
Familiar face	$M = 0.16$ ($SE = 0.04$)	$M = 0.16$ ($SE = 0.04$)	$M = 0.18$ ($SE = 0.05$)	$M = 0.19$ ($SE = 0.5$)
Unfamiliar face	$M = 0.18$ ($SE = 0.06$)	$M = 0.28$ ($SE = 0.06$)	$M = 0.24$ ($SE = 0.06$)	$M = 0.20$ ($SE = 0.06$)

Unfamiliar faces thus elicited significantly more blinks than familiar faces (particularly if they were also young and female), and blinks decreased during the experiment, but not linearly. Young participants made few blinks throughout, but the

number of blinks that old participants made decreased considerably over trial blocks. However, there were very few differences in absolute terms.

2.4. Discussion

This study aimed to investigate the time-course of face learning and whether it could be measured indirectly with physiological responses. We also recorded decision responses that provided an accuracy score, implying that explicit face learning had occurred; and RTs, which have been used to test cognitive processing in many experiments (e.g. Keyes & Zalicks, 2016). These two behavioural responses therefore provided information about the time-course of face learning against which we could measure the physiological responses.

We first examined decision responses, as a conventional measure of face learning. We predicted that accuracy would increase gradually as faces were learnt, in line with Tong and Nakayama, (1999); Kosaka et al., (2003); and Pilz et al. (2009). This was confirmed: overall, accuracy improved gradually from 82% to 90%, for both familiar and unfamiliar faces. The decision responses allowed us to make inferences about explicit face learning: overall most of it had occurred by the second trial block, but improvements were also significant between the fourth and fifth presentations. However, this process may have been affected by the high scores of some participants, who had scores at ceiling for most of the experiment. Indeed, the accuracy rates were higher than those obtained by Bruce et al. (2001), possibly because they used monochrome CCTV videos, while we used high quality colour videos of a talking head which contained better information about each face.

The pattern also suggested that when presented with familiar and unfamiliar faces in an experimental paradigm, unfamiliar face classification improved at a similar rate to familiar face classification. This may have occurred because the unfamiliar faces became easier to categorise as unfamiliar as participants gained more information about the familiar faces to which they had to compare the unfamiliar faces (this is discussed further when discussing pupillary responses).

However, there was no difference in performance between familiar and unfamiliar face processing overall. There were also no significant differences in accuracy between the participant age groups and no overall interactions between *face age* and *participant age*. The data suggest that our face classification task was too easy for it to be able to tease apart age biases. It may be that differences in performance accuracy would emerge between age groups when combining unfamiliar and other-age face processing in a more difficult experimental task. Finally, there were surprising results when looking at gender bias, as when faces were male and young, male participants were less accurate at processing them than when processing other face types, and male participants were less accurate than female participants overall.

This conflicts with the work of Wright and Sladden (2003), who found an own gender bias (OGB): both males and females are better at recognising the faces of people from their own gender. However, it may be accounted for in part by evidence of an own-gender memory bias that is only present in women (Lovén, Herlitz, & Rehnman, 2011). In the present experiment, males were less accurate than women, particularly when male participants processed young male faces. When these participants were also old and the faces were also unfamiliar, accuracy was at 75%. It is possible that a female-only gender bias advantages females when processing female faces even when they are other-age *and*

unfamiliar. However, as males do not have this gender bias, combining the difficult additional tasks of unfamiliar and other-age face processing may mean that accuracy is compromised, particularly in older men.

This is supported in part by the work of Lovén et al. (2012), who tested the combined biases of race and gender. They found that when faces were own-race, more female faces were remembered than male faces, and the scores were higher than when processing male faces or other-race faces, particularly by female participants. However, when faces were other-race, female faces were only remembered better than male faces in female participants (this was the highest score for other-race faces). In male participants, there was no difference in memory for male and female other-race faces. In other words, it appeared that the gender bias found in females overrode the additional difficulties of other-race face processing fairly successfully, indicating that a gender bias can improve accuracy in own-age or own-race faces, and reduce the difficulties found when processing other-age or other-race faces.

We also measured RTs, and our prediction was confirmed that they would be longer for unfamiliar faces and that they would decrease as previously-seen faces became more familiar. Younger participants responded faster than older participants, and their RTs did not decrease as much as those of old participants. Overall, young faces were looked at for longer than old faces; this was because young participants reacted similarly quickly to the faces of different ages, but old participants reacted more slowly to young faces. This could be explained by an asymmetrical OAB, which suggests that people are better at recognising own-age faces better than other-age faces (Anastasi & Rhodes, 2005; see Rhodes & Anastasi, 2012, for a meta-analysis). In this study, there was no difference in accuracy between age groups, but the RT data suggests that old participants took longer

to reach the same degree of accuracy when looking at young faces than young participants did when looking at old faces. This suggests that the OAB can account for the asymmetrical effects, and either that the task was too easy to detect an OAB in young participants or that it was absent in this group. Finally, male participants reacted more slowly than females, as males were particularly slow at processing young faces. When faces were unfamiliar, young and female, males took even longer to process them. This finding lends support to the asymmetrical gender biases described above (Lovén et al., 2011; & Lovén et al., 2012).

The results suggest that RTs provide an indirect index of familiar and unfamiliar face processing. However, RTs continued to decrease significantly and linearly throughout the experiment, whereas improvements in accuracy dwindled after the second trial block, suggesting one of two things: either that RTs are inadequate to index real-time face learning when compared to decision responses, or that they are more sensitive to face learning than decision responses as they continue to diminish after decision responses have plateaued. Indeed, it was probably the case that they were more sensitive than decision responses, as there were many more effects and interactions in the RT data than there were in the decision responses data.

(Note: RTs cannot be used to make inferences about participants' face processing accuracy when participants were categorised by age: old participants took significantly longer to respond, but were only marginally less accurate. It appeared that old participants required more time to achieve a similar level of accuracy as young participants. RTs were therefore helpful for indexing the different speeds at which different face types were processed, but unhelpful in indexing between-participant accuracy.)

Our focus was on pupillary responses, which also appeared to be good indices of differences between familiar and unfamiliar face processing: pupil sizes were larger when viewing unfamiliar faces, suggesting that unfamiliar faces required greater mental effort to process than familiar faces. Pupillary responses were also good indirect indices of explicit face learning: pupil sizes gradually reduced with the steepest decrease in the first trial blocks when looking at familiar and unfamiliar faces, mirroring the improvements in accuracy fairly well.

However, unlike accuracy, where there was no effect of *familiarity* or interaction between *familiarity* and *trial block,* the pupillary data had both: pupil sizes were larger when looking at unfamiliar faces and decreased more quickly in the first trial block when looking at the unfamiliar faces (as can be seen in the steeper trajectories in figs. 12 & 13). This suggests that making decisions about unfamiliar faces required more mental effort than making decisions about familiar faces, something which was more noticeable at the start of the experiment. These findings might have occurred because it was less effortful to match a familiar face to a fragmented representation (positive information) than to reject an unfamiliar face based on fragmented information about other faces (negative information), although the degree of effort had no bearing on accuracy.

One explanation for the differences between how the accuracy and pupillary data changed over successive trial blocks is that perhaps there was a response bias, i.e. a tendency to see all faces as "unfamiliar", which worked in favour of detecting unfamiliar faces but worked against identifying familiar ones. For instance, Jenkins, White, Van Montfort and Burton (2011) showed participants 40 faces of two Dutch celebrities: people who were unfamiliar with the celebrities tended to categorise the faces as being many different individuals, whereas people who were familiar with the celebrities correctly

appreciated that the images contained different views of just two individuals. Similarly, Bindemann and Sandford (2011) tested people's ability to match the correct person (from 30 photographs of different people) to three ID cards (each containing a different photo of that person). They found that people tended to think that the ID cards displayed photos of different people, also suggesting that people have a tendency to regard different views of the same unfamiliar face as belonging to different individuals.

Pupil sizes were larger for old participants than young participants, and old participants had a smaller reduction in pupil size than young participants. This might be thought to indicate that the larger pupils seen in old participants were due to them requiring greater mental effort to achieve the same level of accuracy as young participants, lending further support to an asymmetrical age bias (Anastasi & Rhodes, 2005; see Rhodes & Anastasi, 2012, for a meta-analysis). However, first, the 'larger' pupil sizes between participants are misleading for indexing cognitive load between participants, as they were calculated separately for each participant to represent their relative pupillary changes, and were not absolute values of pupil size. Also, the smaller pupil size reduction could be an effect of age-related changes in autonomic function that result in smaller pupil size changes in older people (Bitsios, Prettyman, & Szabadi, 1996).

The lack of an interaction between *face age* and *participant age* suggests that the mental effort required to process the other-age and own-age faces was no different between age groups. This contrasts with our expectations based on the OAB, although previous research tends to focus on accuracy rather than mental effort (see Rhodes & Anastasi, 2012). We expected other-age faces to be associated with greater cognitive load, and consequently to elicit larger pupil sizes. The task may have been too easy to tease apart direct effects of age bias. However, there was a three-way interaction between

trial block, *face gender* and *participant gender* that showed that pupil sizes were largest for male participants, and that pupil sizes reduced most steadily when female participants looked at female faces, indicating that mental effort was smallest and decreased the most when females learnt female faces, lending further support for an asymmetrical gender bias (Lovén et al., 2011; & Lovén et al., 2012). However, pupils were similar sizes when looking at male faces, smallest when looking at young female faces, and largest when looking at old female faces, indicating that face age can moderate the effects of gender.

The pupillary responses thus provide support for the idea that unfamiliar faces induce greater cognitive load than familiar faces. They also support the inferences from the decision responses, suggesting that cognitive load decreases as faces are learnt. Finally, they indicate that the mental effort of correctly classifying unfamiliar faces decreased as the familiar faces became more familiar and the negative information derived from them increased. Nevertheless, the unfamiliar faces elicited larger pupil sizes (than familiar faces) at each trial block of the experiment, suggesting that classifying an unfamiliar face as unfamiliar at each stage of the learning process is comparatively harder than that of classifying a familiar face as familiar. However, while pupillary responses can index relative changes of within-participant face processing, like RTs, they cannot be used to make inferences about face processing accuracy between participants.

Fixations have seldom been used to measure cognitive load (Ikehara & Crosby, 2005), but they have been studied in face processing research, mainly to investigate where people look when processing faces (e.g. Barton et al., 2006; Van Belle, 2010). However, results are conflicting. Ikehara and Crosby found that the number of eye movements decreased as the task became harder, while the present research found the opposite. It remains unclear whether familiar and unfamiliar faces elicit different numbers of

fixations: Van Belle found no difference in the number of fixations between face types, whereas Barton et al. found that unfamiliar faces elicited more fixations than familiar faces.

The present research also found that unfamiliar faces elicited more fixations than familiar faces, suggesting that people need to look at or double-check more points on an unfamiliar face to process it. The number of fixations also decreased during the experiment indicating that participants needed fewer fixations to classify faces (as familiar or unfamiliar), as the familiar faces became more familiar. This indicates that the number of fixations is also a good indirect index of face learning.

We found that older participants made more fixations than younger participants, suggesting that they needed to look at or double-check more points on the faces than young participants. The number of fixations that old participants also decreased more dramatically than those of young participants. Old participants made almost double the number of fixations when looking at unfamiliar faces than did young participants (young participant: $M = 3.34$, $SE = 0.34$; old participant: $M = 6.39$, $SE = 0.39$), and double the number of fixations when looking at young faces than did young participants (young face: $M = 3.28$, $SE = 0.31$; old face: $M = 6.05$, $SE = 0.27$). However, the effect was not symmetrical. This suggests that old participants might treat other-age faces as unfamiliar, partially supporting the OAB (Anastasi & Rhodes, 2005). Partial OAB and gender biases were also reflected in some three-way interactions: young unfamiliar female faces elicited the most fixations, almost two more fixations than familiar old male faces (which elicited the fewest); and while all participants made the fewest fixations when looking at familiar old faces and the most while looking at unfamiliar young faces, this was most evident in male participants who made almost 1.5 times more fixations while looking at unfamiliar

young faces than did females. This lends further support to the idea that processing faces that combine the more difficult tasks of other-age, other-gender, and unfamiliar face processing is more effortful in older males.

Overall, the results suggest that fixations are useful in indexing familiar and unfamiliar face processing and face learning. However, as the average number of fixations per image was very low in this experiment, with some participants only requiring one per image, it was only when faces became harder to process by combining unfamiliarity, other-age and/or other-gender, that age and gender effects were detected, and this only occurred in old participants.

Finally, blinks have been associated with cognitive load (Martins & Carvalho, 2015). We found that participants also made more blinks when looking at unfamiliar faces, apart from when they looked at unfamiliar, young, male faces. When participants looked at these they blinked a similar number of (fewer) times to when they looked at familiar faces. We also found that the number of blinks old participants made decreased during the experiment. Old participants blinked more when looking at familiar old faces than when looking at familiar young faces; and old participants blinked more when looking at old male faces than when they looked at old female faces. These findings lend further support to the idea that combining the more difficult tasks of processing unfamiliar faces, other-age faces, and other-gender faces affects older people more than younger people. While blinks seemed to index familiar and unfamiliar processing in this experiment, they were less effective at indexing face learning, particularly in young participants. Like fixations, the average number of blinks per image was very low in this experiment, with many images not producing any blinks at all, so this measure was inadequate to index fluctuations in cognitive load associated with learning.

Thus, pupillary responses appear to be the most satisfactory physiological indices of familiar and unfamiliar face processing and real-time face learning, as they are not limited by the floor effects seen in the fixation and blink data, at least in this type of experiment. The pupillary data suggest that the early stages of face learning are discernible during a single experimental session, and that it occurs gradually (with the greatest changes occurring at the beginning of the experiment then dwindling). They also show that it becomes easier to decide that an unfamiliar face has not been seen before as the faces that have been seen previously become more familiar. It seems that cognitive load can account for pupillary changes during face learning: the pupillary data indicate that processing unfamiliar faces involves a greater cognitive load than processing familiar faces, and that cognitive load decreases gradually as faces become more familiar. These effects occur in old and young participants, although some differences are found between the two age groups. While they cannot replace conventional decision responses as measures of learning, they provide important indications of the nuances of familiar and unfamiliar face processing and face learning associated with age and gender, and combined with the measures described here, provide partial support for age and gender biases.

However, there were potential issues with the physiological responses that we tested. While there was only a marginal effect of age on *accuracy*, all three physiological measures (and RTs) found significant asymmetrical age differences. This could indicate that old participants require more mental effort to be accurate than young participants, but these differences could also be artefacts of *physical* ageing. For instance, older participants are more likely to suffer from dry eyes (the National Eye Institute, 2017), so they might need to blink more to maintain moisture, rather than blinking more because

of increased mental effort. Age-related changes are also found in the autonomic function, evidenced by reduced pupillary fluctuations (Bitsios et al., 1996).

Therefore, to test for physiological other-type face effects while controlling for physical effects of ageing, we conducted a second experiment that tested two groups of young participants of different races, evaluating the other race effect (ORE). This is similar to the OAB, but shows that own-race faces are easier to classify and learn than other-race faces (e.g. Meissner & Brigham, 2001; Michel, Rossion, Han, Chung, & Caldara, 2006; Meissner, Susa, & Ross, 2013), even when the memory demands of face recognition have been minimised, such as in face matching tasks (e.g. Megreya, White, & Burton, 2011).

Experiment 2

2.5. Methods

2.5.1. Participants

Thirty-two participants with normal or corrected to normal vision were recruited via the university, in exchange for cash or course credits. There were sixteen males and sixteen females, all aged between 18 and 27. Half were Caucasian, and half were far east Asian (eight males and eight females in each group). All participated in both experimental sets. The Caucasian participants had also participated in Experiment 1, and acted as controls in this experiment. In Experiment 1, sixteen of these had participated in the main experiment while six had participated in the pilot. Pilot participants were included in the main experiment. As it was only possible to recruit sixteen Asians in Experiment 2, we

only included the sixteen participants who had participated in the main experiment in Experiment 1.

2.5.2. Apparatus and Materials

These were the same as in Experiment 1, except that the 'old' stimuli were replaced with 'Asian' stimuli which were matched to the 'Caucasian' stimuli in terms of age. These were taken from either VidTIMIT (2009), and the FEI Face Database (2006), or the CUHK Face Dataset (CUHK, 2009).

2.5.3. Design

This study also used a mixed design: repeated measures on *trial block* (with six trial blocks: 1, 2, 3, 4, 5 or 6), *familiarity* (with two familiarity types: familiar and unfamiliar), *face race* (with two face types: Caucasian and Asian), *face gender* (with two face types: male and female), and independent measures on *participant race* (with two race groups: Caucasian and Asian) and *participant gender* (with two genders: male and female). The dependent variables were the same as in Experiment 1.

2.5.4. Procedure

This was the same as in Experiment 1. The procedure was repeated four times, once for each face type (Caucasian male, Caucasian female, Asian male and Asian female).

2.6. Results

In this experiment, we created two graphs for each DV that expressed the data we were most interested in investigating: trial block, participant race, face race, and

familiarity. Therefore, for this experiment, each section has one graph presenting data from *familiar* faces and one presenting data from *unfamiliar* faces over the six sequential trial blocks, as a function of participant race. Any further interactions are presented in additional graphs.

2.6.1. Decision responses (accuracy)

A Mixed ANOVA was performed to compare accuracy while viewing familiar and unfamiliar faces.

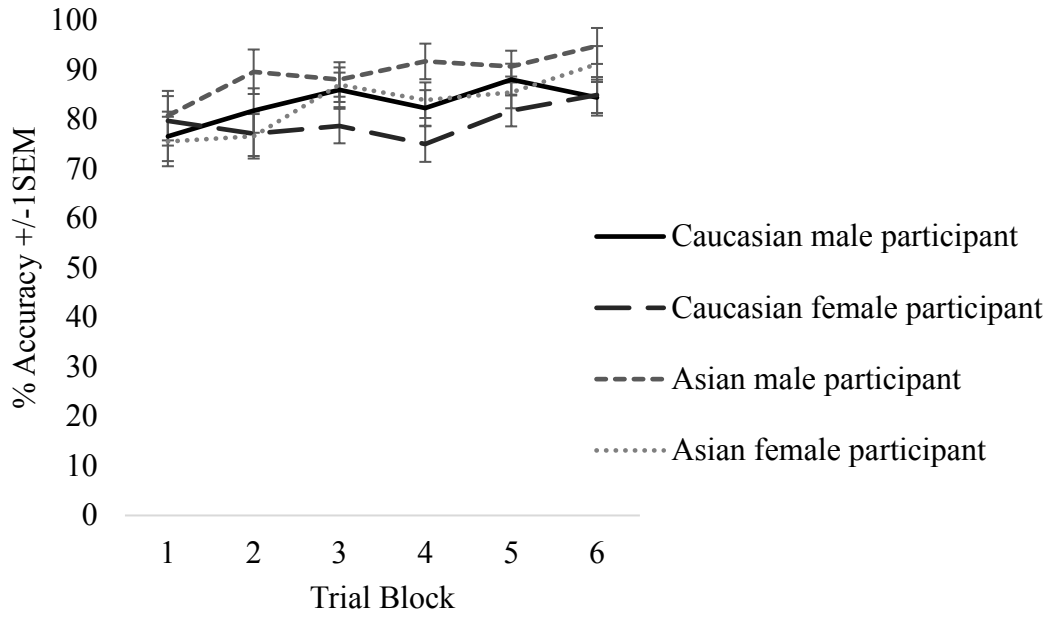


Fig. 22. Mean accuracy for *familiar* faces over six sequential trial blocks, as a function of participant race and gender.

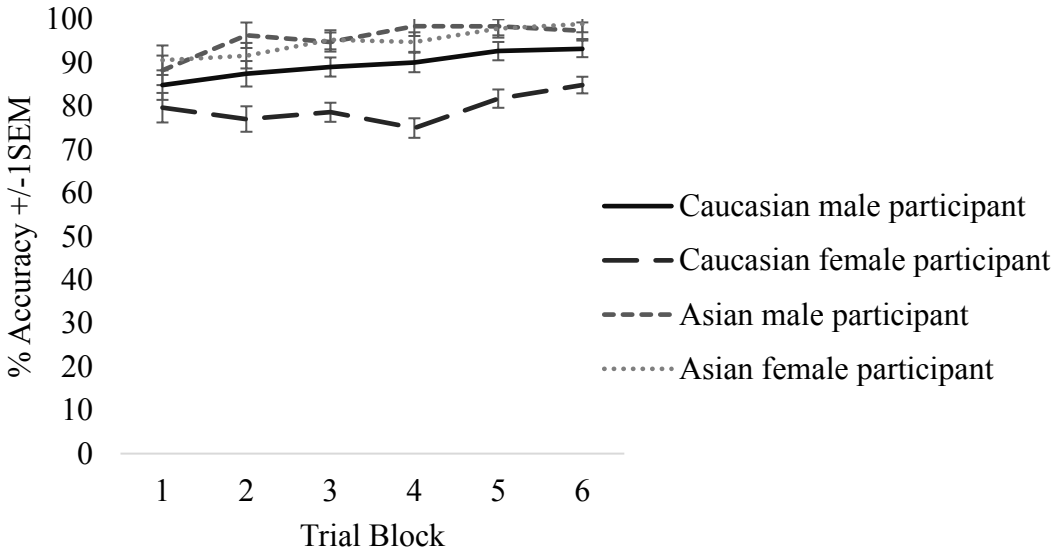


Fig. 23. Mean accuracy for *unfamiliar* faces over six sequential trial blocks, as a function of participant race and gender.

As shown in figs. 22 and 23, the results show that there were significant effects of *trial block*, F (5, 140) = 15.74, $p < .001$, $r = .33$, $\eta^2 = .36$. Planned contrasts revealed that Caucasian participants showed significant improvements in accuracy only between the fourth and fifth trial blocks ($p < .05$), while in Asian participants showed significant improvements between the first and second ($p = .01$), the fifth and sixth ($p < .05$). No other comparisons reached significance.

There was also a significant effect of *familiarity*, F (1, 28) = 26.80, $p < .001$, $r = .70$, $\eta^2 = .49$ (familiar face: $M = 83.77$, $SE = 1.55$; unfamiliar face: $M = 93.04$, $SE = 0.82$), and *participant race*, F (1, 28) = 7,43, $p = .01$, $r = .46$ (Asian participant: $M = 90.74$, $SE = 1.21$; Caucasian participant: $M = 86.07$, $SE = 1.21$), and a significant interaction between *trial block* and *participant race*, F (5, 140) = 2.26, $p = .05$, $\eta^2 = .30$. There was no significant effect of *participant gender*, F (1, 28) = 1.32, $p = .26$.

Results also found effects of *face race*, F (1, 28) = 9.57, $p = .01$, $r = .50$, $\eta^2 = .26$ (Asian face: $M = 86.16$, $SE = 1.18$; Caucasian face: $M = 90.65$, $SE = 1.06$), and *face gender*, F (1, 28) = 22.74, $p < .001$, $r = .67$, $\eta^2 = .45$ (male face: $M = 91.69$, $SE = 0.97$; female face: $M = 85.12$, $SE = 1.21$). Finally, there was an interaction between *face race* and *participant race*, F (1, 28) = 19.06, $p < .001$, $\eta^2 = .41$ (Asian participant, Asian face: $M = 91.66$, $SE = 1.66$; Asian participant, Caucasian face: $M = 89.81$, $SE = 1.50$; Caucasian participant, Asian face: $M = 80.66$, $SE = 1.66$; Caucasian participant, Caucasian face: $M = 91.48$, $SE = 1.50$), but no other interactions reached significance.

This indicated that participants were better at making decisions about unfamiliar faces than familiar faces, and that accuracy rates for both familiarity conditions increased significantly across trial blocks. As the familiar faces became more familiar, the ability to classify them as familiar increased, as did the ability to classify unfamiliar faces as

unfamiliar. The results also indicate that participants were better at making familiarity decisions about Caucasian faces than Asian faces.

The interaction between *trial block* and *participant race* indicates that improvement in accuracy for Caucasian participants was slower and less linear than for Asian participants, and the interaction between *face race* and *participant race* demonstrated that there was an asymmetrical ORE that was present for Caucasian participants, but not for Asian participants.

2.6.2. Reaction Times

A similar analysis was performed to analyse the reaction time data. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 39.05, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .66$).

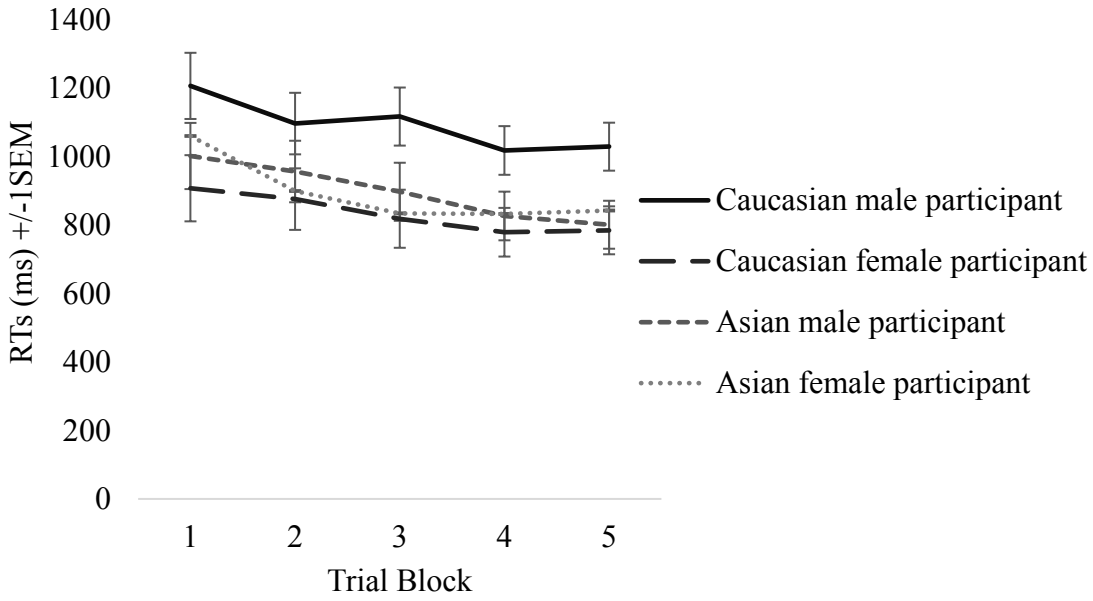


Fig. 24. Mean RT for *familiar* faces over six sequential trial blocks, as a function of participant race and gender.

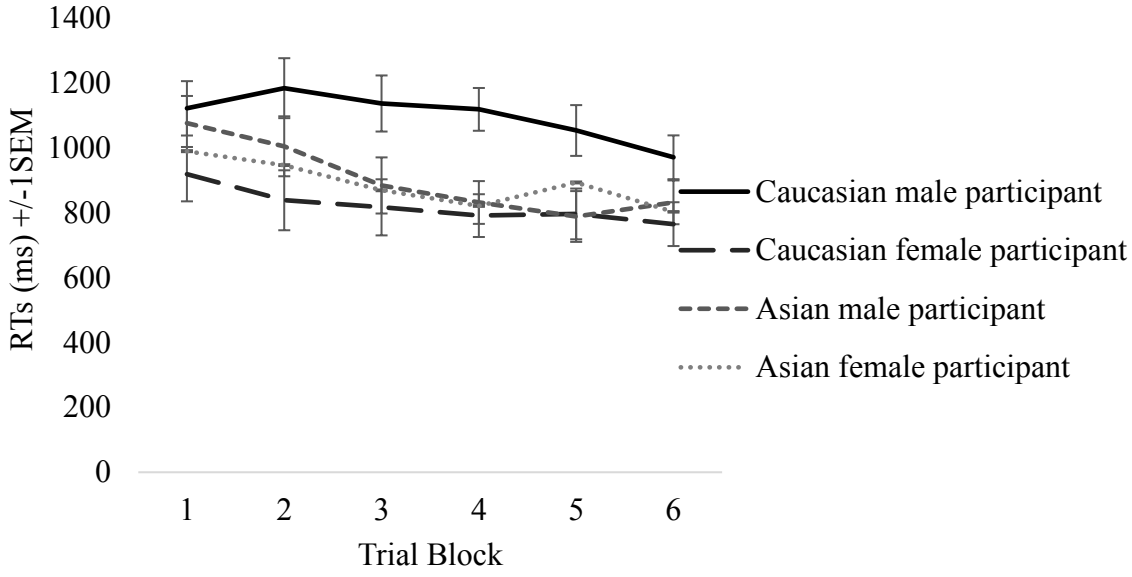


Fig. 25. Mean RT for *unfamiliar* faces over six sequential trial blocks, as a function of participant race and gender.

As shown in figs. 24 and 25, there was a significant effect of *trial block*, F (3.30, 92.31) = 24.01, $p < .001$, $r = .45$, $\eta^2 = .46$; where planned contrasts revealed that there were only significant reductions in RT between the fifth and sixth trial blocks in Caucasian participants (p =.01), while in Asian participants there were significant reductions in RT between the first and second (p =.01), the second and the third (p =.01), and the third and fourth trial blocks (p =.03). No other comparisons reached significance,

As seen in fig. 26, there was also a three-way interaction between *trial block*, *face gender* and *participant gender*, F (3.93, 110.00) = 3.25, p = .02, η^2 = .39: RTs were slowest when males looked at female faces and fastest when females looked at male faces.

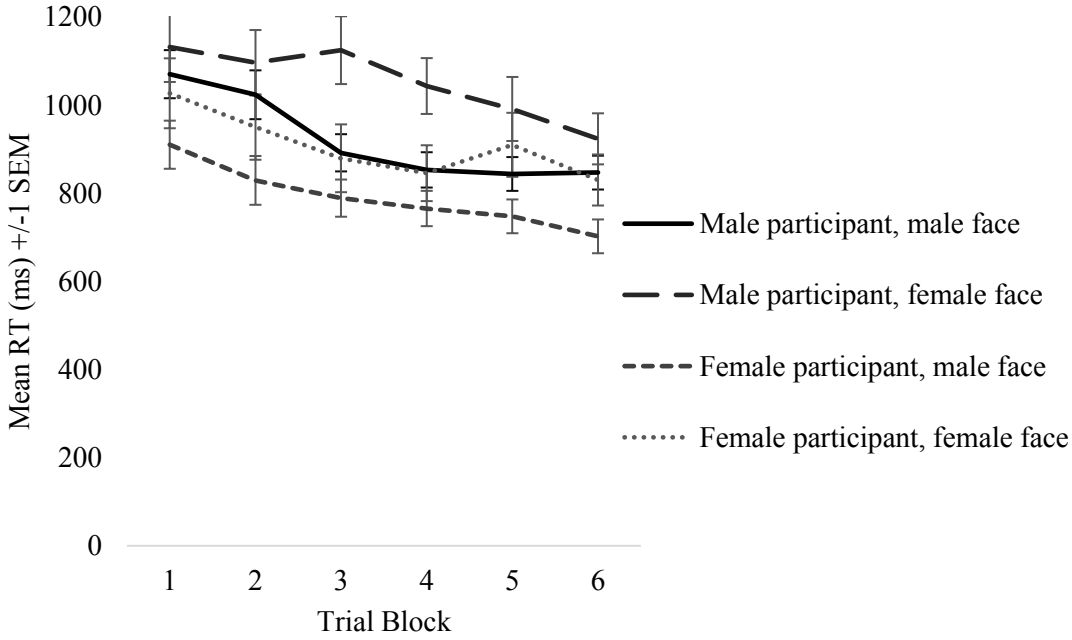


Fig. 26. Mean RT in *all* participants over six sequential trial blocks, as a function of face gender and participant gender.

There were no effects of *familiarity*, F (1, 28) = 0.57, p = .46, *participant race, F* (1, 28) = 0.95, p = .34, or *face race F* (1, 28) = 1.94, p = .18. However, results showed

that there was a significant effect of *face gender*, F (1, 28) = 15.20, p = .01, r = .59, η^2 = .35 (male face: M = 856.14, SE = 27.51; female face: M = 979.50, SE = 46.35).

As seen in fig 27, there was also an interaction between *familiarity* and *face gender*, F (1, 28) = 7.06, p = .01, η^2 = .20 (familiar male face: M = 828.45, SE = 27.32; familiar female face: M = 990.54, SE = 52.29; unfamiliar male face: M = 883.83, SE = 27.51; unfamiliar female face: M = 968.47, SE = 45.64).

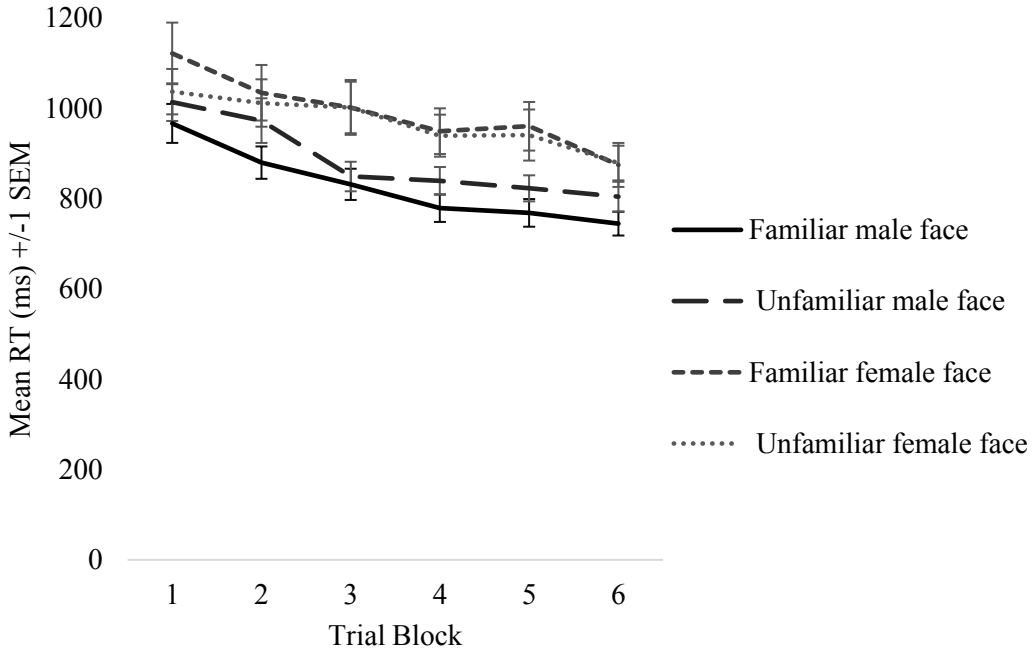


Fig. 27. Mean RT in *all* participants over six sequential trial blocks, as a function of face gender and familiarity.

The results indicate that RTs decreased linearly, and that female faces elicited longer RTs than male faces, particularly when the male faces were familiar. Also, while participants looked at faces of their own gender a similar amount of time, males looked for longest time at female faces, and females looked at male faces for the shortest time.

2.6.3. Pupillary responses

A similar analysis was performed to analyse pupil sizes while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 79.63, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .44$).

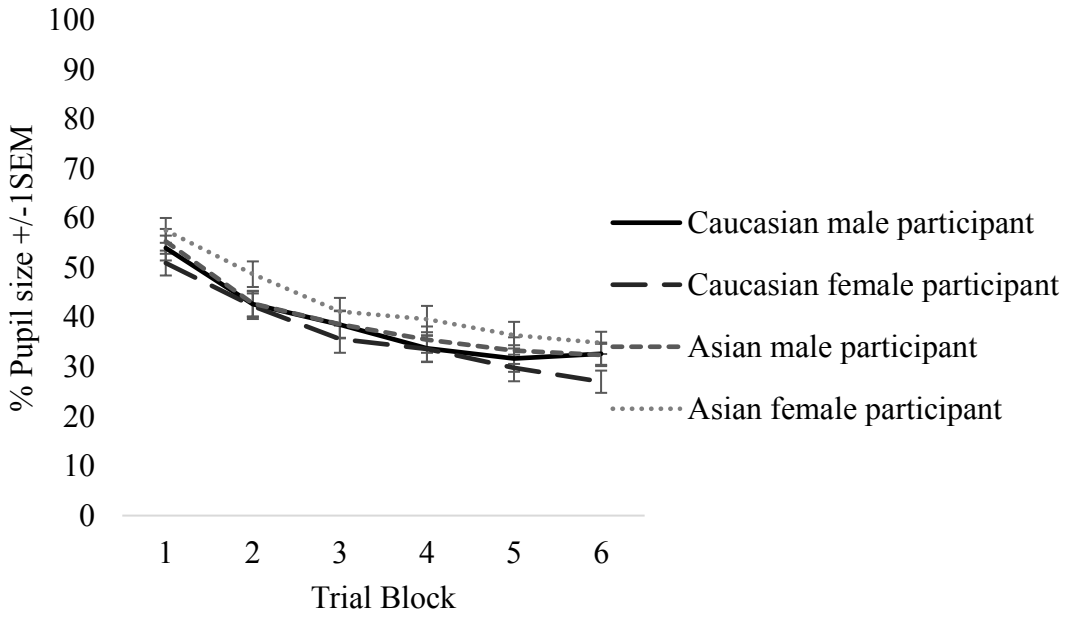


Fig. 28. Mean pupil size for *familiar* faces over six sequential trial blocks, as a function of participant race and gender.

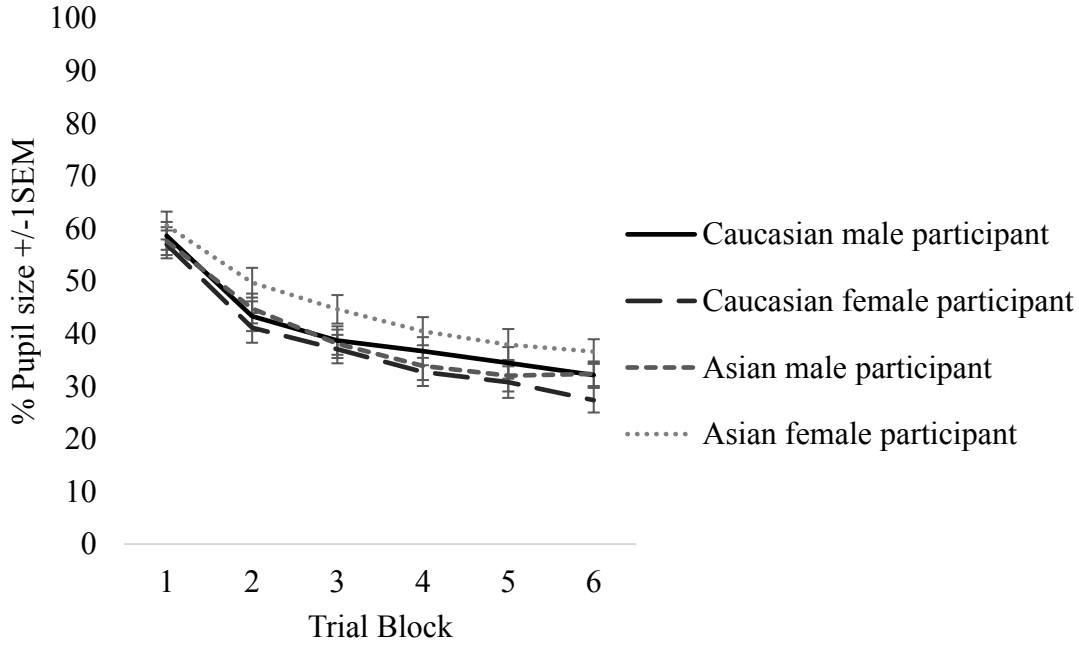


Fig. 29. Mean pupil size for *unfamiliar* faces over six sequential trial blocks, as a function of participant race and gender.

As shown in figs. 28 and 29, the results show that there were significant effects of *trial block*, F (2.18, 61.01) = 219.52, $p < .001$, $r = .88$, $\eta^2 = .89$; where planned contrasts revealed that the first three trial blocks elicited a significantly smaller pupil size than the preceding one (all ps < .001) in both Asian and Caucasian participants. In Caucasian participants, there was also a significantly smaller pupil size in the fifth compared with the fourth trial block ($p = .01$) and in the sixth compared with the fifth ($p = .04$). In Asian participants, there was also a significantly smaller pupil size in the fifth compared with the fourth ($p < .001$), but pupil sizes were no smaller in the sixth compared with the fifth ($p = .41$).

There were also significant effects of *familiarity*, F (1, 28) = 6.80, $p = .01$, $r = .44$, $\eta^2 = .20$ (familiar face: $M = 39.56$, $SE = 1.15$; unfamiliar face: $M = 40.79$, $SE = 1.18$), and an interaction between *trial block* and *familiarity, F* (4.33, 121.28) = 5.31, $p < .001$, $\eta^2 = .04$. Results also showed an effect of *face race*, F (1, 28) = 12.73, $p < .001$, $r = .56$, $\eta^2 = .21$ (Asian face: $M = 43.65$, $SE = 1.43$; Caucasian face: $M = 36.71$, $SE = 1.57$), $r = .56$.

However, there were no effects of *face gender*, F (1, 28) = 0.61, $p = .44$, (male face: $M = 39.75$, $SE = 1.44$; female face: $M = 40.61$, $SE = 1.38$), *participant gender*, F (1, 28) = 0.13, $p = .72$, (male participant: $M = 39.77$, $SE = 1.61$; female participant: $M = 40.59$, $SE = 1.561$), or *participant race*, F (1, 28) = 2.28, $p = .14$, (Asian participant: $M = 41.90$, $SE = 1.61$; Caucasian participant: $M = 38.46$, $SE = 1.61$). No other interactions reached significance.

The results indicated that unfamiliar faces elicited significantly larger pupil sizes than familiar faces, and that pupil sizes decreased with each successive trial block. The interaction between *trial block* and *familiarity* indicated that the familiar and unfamiliar faces elicited different time-courses of pupil reduction, with a steeper initial decrease for

unfamiliar faces compared to familiar faces. The results also revealed that Asian faces elicited larger pupil sizes throughout the experiment than Caucasian faces, although the pupil size reduction trajectory was no different between *face races*, and the lack of interaction between *participant race* and *face race* indicated that there was no discernible ORE on pupil size.

2.6.4. Fixations

A similar analysis was performed to analyse the number of fixations while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 49.40, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .61$).

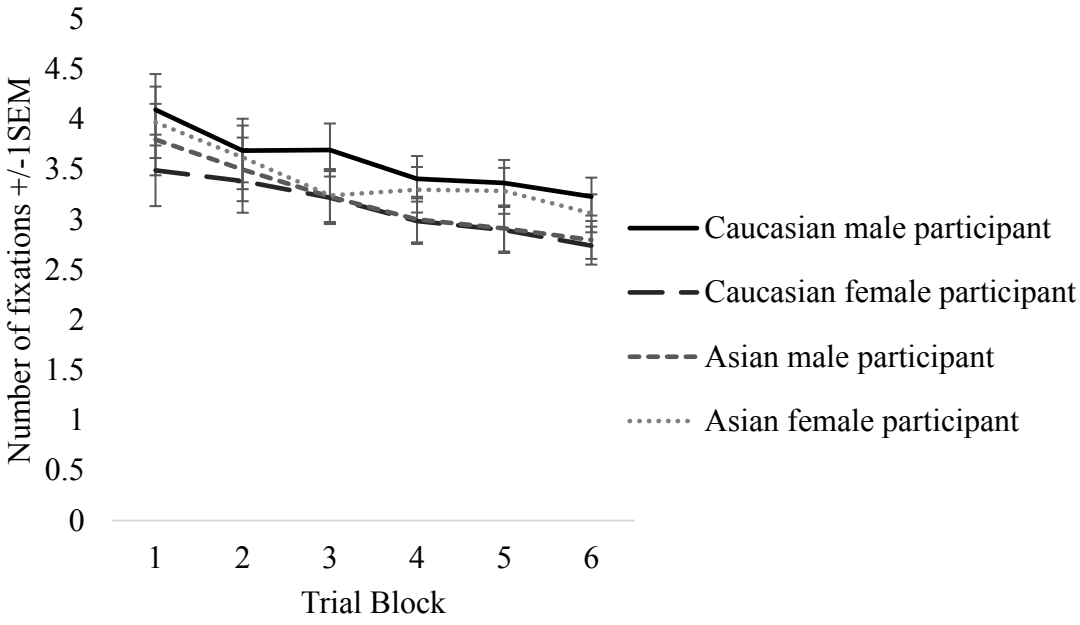


Fig. 30. Mean number of fixations for *familiar* faces over six sequential trial blocks, as a function of participant race and gender.

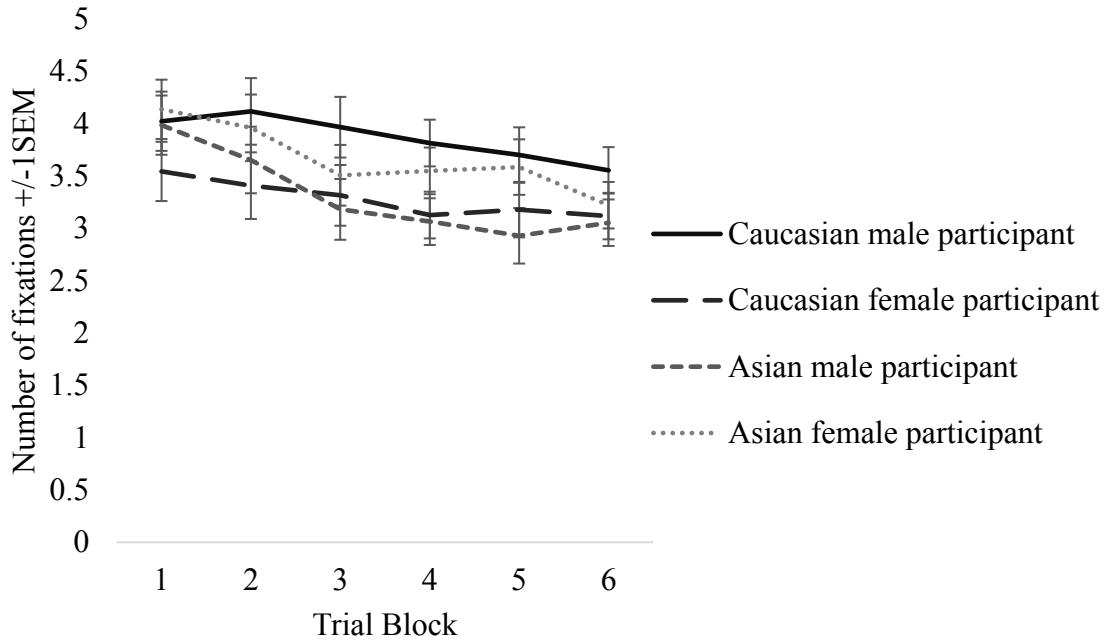


Fig. 31. Mean number of fixations for *unfamiliar* faces over six sequential trial blocks, as a function of participant race and gender.

As shown in figs. 30 and 31, there were significant effects of *trial block*, F (3.03, 84.71) = 34.70, $p < .001$, $r = .54$, $\eta^2 = .55$; where planned contrasts revealed that the fifth trial block elicited significantly fewer fixations than the fourth trial block in Caucasians ($p = .03$), but in Asians, the second elicited fewer than the first ($p = .01$), the third elicited fewer than the second ($p < .001$), and the sixth elicited significantly fewer than the fifth ($p < .05$).

There was also a significant effect of *familiarity*, F (1, 28) = 7.24, $p = .01$, $r = .45$, $\eta^2 = .21$ (familiar face: $M = 3.33$, $SE = 0.12$; unfamiliar face: $M = 3.53$, $SE = 0.12$), $r = .45$, *face race*, F (1, 28) = 4.56, $p = .04$, $r = .37$ (Asian face: $M = 3.58$, $SE = 0.15$; Caucasian face: $M = 3.28$, $SE = 0.12$), and *face gender*, F (1, 28) = 22.40, $p < .001$, $r = .67$, $\eta^2 = .44$ (male face: $M = 3.20$, $SE = 0.10$; female face: $M = 3.65$, $SE = 0.15$).

However, there were no significant effects of *participant gender*, F (1, 28) = 0.27, $p = .61$, or *participant race*, F (1, 28) = 0.24, $p = .79$. There were also no significant interactions.

The results indicate that fixations decreased (linearly) during the experiment, unfamiliar faces elicited significantly more fixations than familiar faces, Asian faces elicited significantly more fixations than Caucasian faces, and female faces elicited significantly more fixations than male faces. However, in *absolute* terms, there was very little change in the number of fixations over time.

2.6.5. Blinks

A similar analysis was performed to analyse the number of blinks while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 43.04, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .59$).

The results show that there were no significant effects. The largest F was for *familiarity*, F (1, 28) = 3.49, $p = .07$ (familiar: $M = 0.14$, $SE = 0.4$; unfamiliar: $M = 0.16$, $SE = 0.5$). Also, in absolute terms, there were very few blinks in response to the faces.

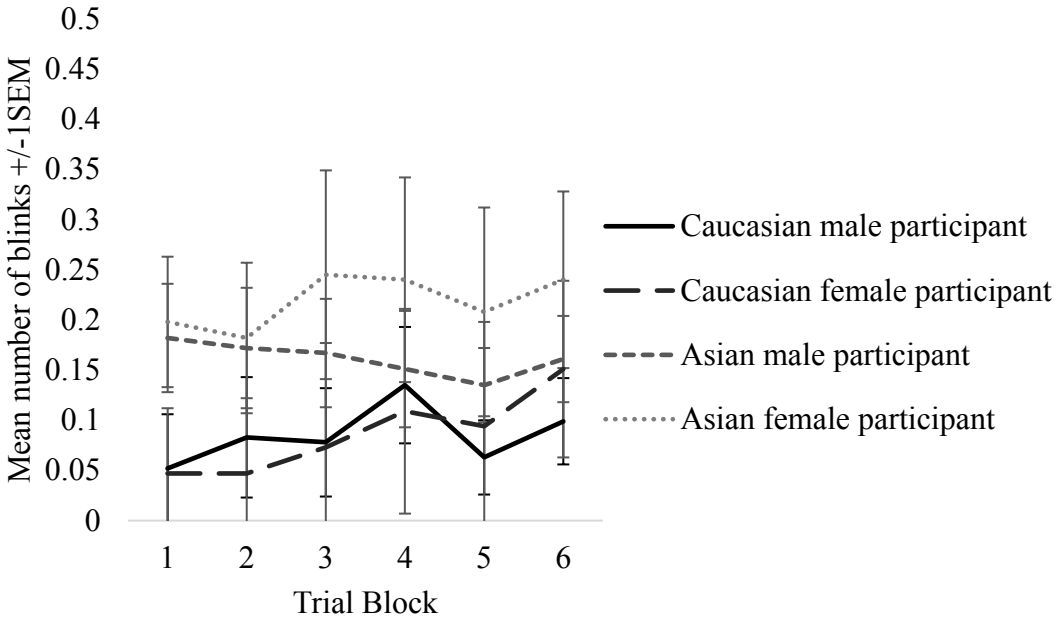


Fig. 32. Mean number of blinks for *familiar* faces over six sequential trial blocks, as a function of participant race and gender.

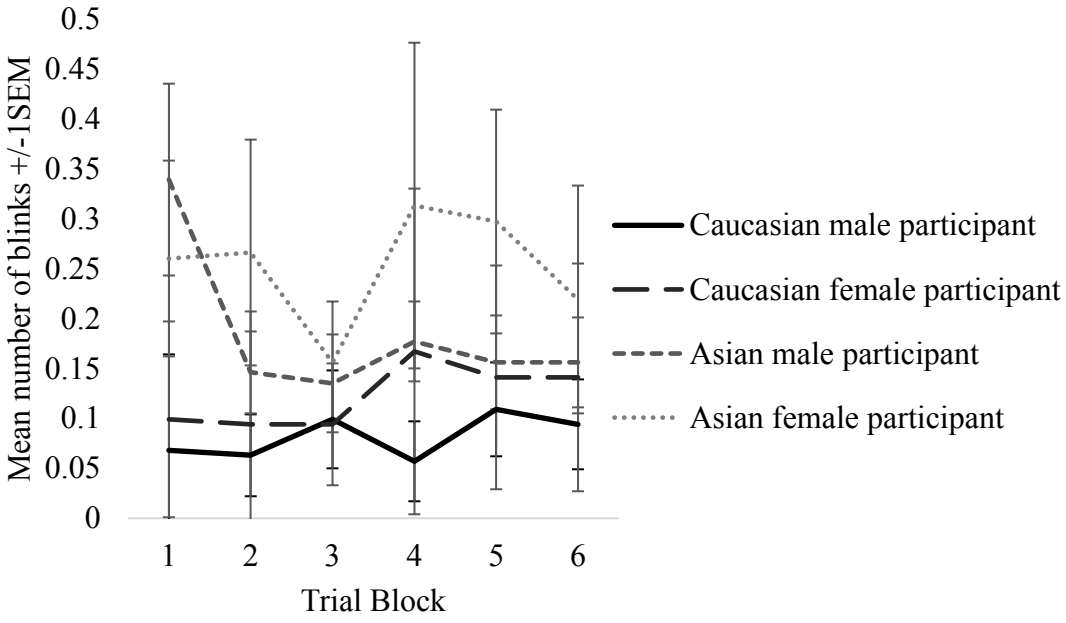


Fig. 33. Mean number of blinks for *unfamiliar* faces over six sequential trial blocks, as a function of participant race and gender.

The results indicate that blinks are not good indices of mental effort while processing different face types, at least not in this type of experiment.

2.7. Discussion

This experiment had two purposes. First, we wanted to evaluate the findings in Experiment 1 by controlling for the physical effects of age on ocular properties such as pupil size. Experiment 1 showed significant physiological and RT differences between age groups, but no significant effects of age on accuracy, suggesting that some of these physiological differences could be artefacts of physical degeneration of autonomic function or tear ducts, rather than arising from differences in cognitive load. Second, we wanted to test whether the ORE accounts for some pupillary responses in participants of different races.

We first tested decision responses. Overall, improvements in accuracy rose during the experiment from 83% to 92%, although most of this had occurred by the third trial block, suggesting that the faces had been learnt sufficiently by this stage. Overall, Asian participants improved more quickly and were more accurate than Caucasian participants. As in Experiment 1, the pattern suggested that as participants improved in responding to the familiar faces, they also improved in responding to the unfamiliar faces. Indeed, in this experiment, participants were more successful at responding to unfamiliar faces than familiar faces throughout, although they improved faster when responding to familiar faces (as can be seen in figs. 23 & 24).

We also found that there were significant effects of *face gender* on accuracy: people were better at processing male faces than female faces, conflicting with previous research on gender biases (Wright & Sladden, 2003; Lovén et al., 2011; Lovén et al.,

2012), but this might have been because the male faces in this experiment were more distinctive. Indeed, while every attempt was made to match stimuli (and stimulus sets) on the bases of race, race, gender and distinctiveness, many participants of both genders and races said that they found the female faces harder. This was generally because they all had 'long hair', which appeared to make it particularly difficult to respond to the Asian female faces.

There were also effects of *face race*: people were better at processing Caucasian faces than Asian faces, and Asian participants were considerably better at processing Caucasian faces than the Caucasian participants were at processing Asian faces, suggesting that the ORE was asymmetrical. This (and the faster learning by Asian participants) can be explained in several ways. The most well documented is the 'contact hypothesis' (Chiroro & Valentine, 1995), and in this experiment the Asian participants had more contact with Caucasians than the Caucasian participants had with Asians because the Asian participants were international students, studying at the same British university as the Caucasian participants. Another explanation is related to linguistic abilities, as it has been found that the ORE is stronger in monolingual than bilingual people (Kandel et al., 2016). The Asian participants in our experiment were not bilingual but they spoke English, which might have helped them to classify the Caucasian faces, but the Caucasians did not speak any Asian languages, putting them at a disadvantage. An alternative explanation is social-importance: people are more motivated to individuate socially-important faces and disregard socially-unimportant ones (Keyes & Zalicks, 2016). As the Asians were studying at a British university they might have had more motivation to individuate Caucasian faces than their Caucasian counterparts to individuate Asian ones. However, there is also a body of work suggesting that the ORE

is stronger in Caucasian participants than other races regardless of moderators such as contact (Meissner & Brigham, 2001).

As in Experiment 1, RTs gradually decreased (linearly) throughout the experiment. However, while there were no discernible effects of *participant race* on this measure, there were asymmetrical effects of gender that were moderated by *familiarity*: participants took longer to respond to female faces than male faces when faces were familiar, an effect which diminished when the faces were unfamiliar. Also, when female participants looked at female faces, RTs decreased linearly until the fifth trial block, after which they stopped decreasing. In contrast, when male participants looked at female faces, RTs did not decrease until the third trial block, after which they reduced linearly. These findings also support an asymmetrical gender bias (Lovén et al., 2011; Lovén et al., 2012).

As in Experiment 1, the pupillary results showed an interaction between *trial block* and *familiarity* that was absent in the accuracy results: there was a steeper initial decrease in pupil size for unfamiliar faces than for familiar faces, suggesting that the mental effort of processing unfamiliar faces decreased more quickly than did the mental effort of processing the familiar faces. There were also different patterns of pupillary changes for faces of different genders (pupil sizes decreased more linearly for female faces), although there was no interaction between *face gender* and *participant gender*, thus not lending much support for a gender bias (Wright & Sladden, 2003; Lovén et al., 2011; Lovén et al., 2012). Pupil sizes were larger when participants looked at Asian faces compared with Caucasian faces. However, there were no interactions between *face race* and *participant race,* as would be needed to conclude that an ORE existed. Again, the larger pupil sizes for Asian faces might have occurred for various reasons: either the

Asian students had more contact with other-race faces than did the Caucasian participants (Chiroro & Valentine, 1995), because the Asian participants spoke English (Kandel et al., 2016), because the Caucasian faces were more socially important (Keyes & Zalicks, 2016), or because the ORE is stronger in Caucasian participants than other races (Meissner & Brigham, 2001).

Like RTs, the number of fixations decreased progressively during the experiment, but in absolute terms, there was very little decrease over time, suggesting that they were not as useful as the RTs or pupillary responses. However, the fixation data revealed that participants made more fixations when looking at unfamiliar faces compared with familiar faces, when looking at Asian faces compared with Caucasian faces, and when looking at female faces compared with male faces. Therefore, there was partial support for an asymmetrical ORE (e.g. Chiroro & Valentine, 1995; Meissner & Brigham, 2001; Keyes & Zalicks, 2016; Kandel et al., 2016), and for an asymmetrical gender bias (Lovén et al., 2011; Lovén et al., 2012).

The blink data revealed no significant effects, suggesting that they were ineffective at indexing fluctuations in cognitive load associated with face learning, race or gender. These results also indicated that the age-related effects seen in Experiment 1 might be attributable to the physical effects of ageing such as dry eyes (the National Eye Institute, 2017).

In short, while decision responses were not always reliable measures of learning, they are widely used to measure it overtly. Experiment 2 indicated that RTs supported decision responses well although they continued to decrease after the faces had apparently been learnt, and they failed to reveal any fluctuations related to race. Of the three physiological responses, pupillary responses were the most successful in supporting

accuracy, and only pupillary responses showed changes over time that paralleled the improvements in accuracy. Furthermore, only the pupillary responses indicated that mental load decreased faster when classifying unfamiliar faces than familiar faces. As described earlier, this supports the idea that the difficulty in classifying unfamiliar faces was greater when the information about the familiar faces to which they were compared was sparse. The pupillary data also support the idea that the mental load in classifying the unfamiliar faces decreased faster than that it did for familiar faces as the experiment progressed, as the pupil sizes decreased more dramatically when looking at unfamiliar faces.

While all the physiological responses indicated that processing unfamiliar faces required more mental effort than familiar faces, only pupillary and fixation data indicated that there were asymmetrical effects of race (e.g. Chiroro & Valentine, 1995; Meissner & Brigham, 2001; Kandel et al., 2016). Also, trends towards asymmetrical gender biases were only supported by the pupillary data, albeit minimally (Wright & Sladden, 2003; Lovén et al., 2011; Lovén et al., 2012).

However, while conducting this experiment, we became aware that the data on the reduction in pupillary responses and the decline in fixations that we had attributed to mental effort might instead have been artefacts of the participants' varying RTs. We had assumed that because all three measures (RTs, pupil sizes and fixations) had decreased during the experiment, they reflected decreasing cognitive load. However, it became apparent that the dwindling number of fixations could have been an artefact of the experimental procedure: if images disappeared more quickly when RTs were faster, this would leave less time to make any fixations. Similarly, as RTs became faster, there was less time for pupillary changes to occur in response to the image. Between each image

there was a drift check (a white screen) that gave the pupil time to 're-set', so the decreasing pupil sizes might have been the result of less time spent looking at the images between the drift checks, as RTs decreased. Therefore, we decided to test whether pupil sizes and fixations decreased even if participants were presented with images for a fixed time. We tested this in a third experiment, where we also tested the effects of viewing time on blinks and accuracy.

Experiment 3

2.8. Methods

2.8.1. Participants

Twenty-four participants with normal or corrected to normal vision were recruited via the university, in exchange for cash or course credits. There were six males and eighteen females, all aged between 18 and 25 ($M = 19.17$, $SD = 1.61$). All were Caucasian.

2.8.2. Apparatus and Materials

These were the same is in experiment 1, other than that we only used the young Caucasian stimuli. These were taken from either VidTIMIT (2009), or the FEI Face Database (2006).

2.8.3. Design

This study also used a mixed design: repeated measures on *trial block* (with six trial blocks: 1, 2, 3, 4, 5 or 6), *familiarity* (with two familiarity types: familiar and unfamiliar), *viewing time* (with two conditions: fixed and variable) and *face gender* (with two face types: male and female), and independent measures on *participant gender* (with

two genders: male and female). The dependent variables were the same as in Experiment 1.

2.8.4. Procedure

This was the same as in Experiment 1. However, participants only saw the young Caucasian stimuli. In the *fixed* viewing time condition, participants were presented with each image for 5 seconds, while in the *variable* reaction-time dependent viewing time condition, the image was replaced by a drift check as soon as participants had responded. The variable viewing time condition was the same as the procedure used in the preceding experiments.

2.9. Results

In this experiment, we created two graphs for each DV that expressed the data we were most interested in investigating: trial block, participant gender, familiarity, and viewing time. Therefore, in this experiment, each section has one graph presenting data from the fixed viewing time condition and one presenting data from the RT viewing time condition over the six sequential trial blocks, both as a function of participant gender and familiarity. Any further interactions are presented in additional graphs.

2.9.1. Decision responses (accuracy)

Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 33.08, p = .01. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .57).

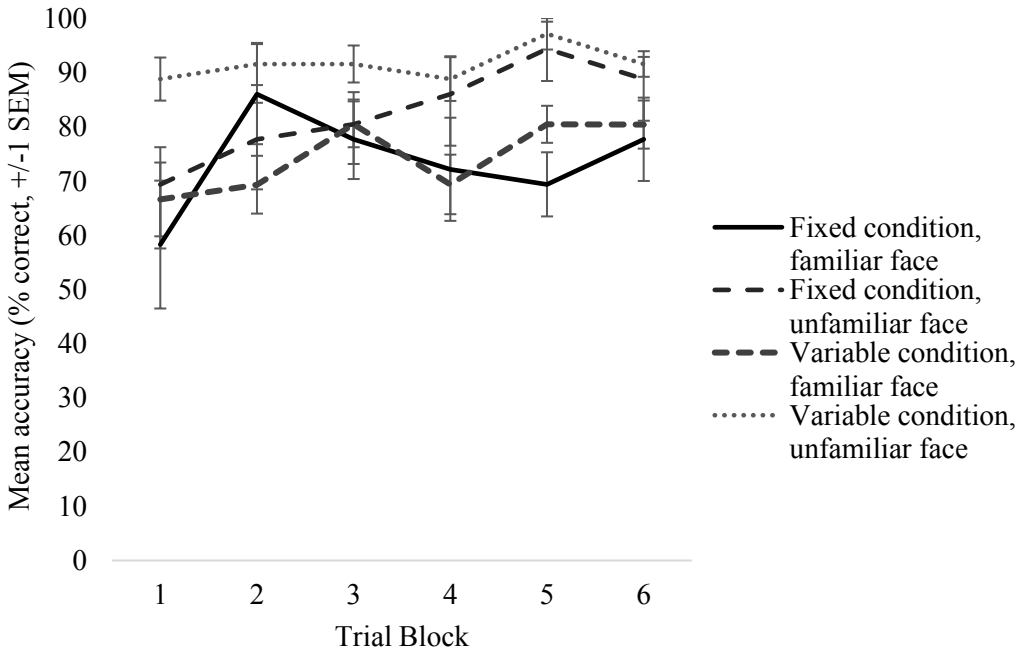


Fig. 34. Mean accuracy in *male* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

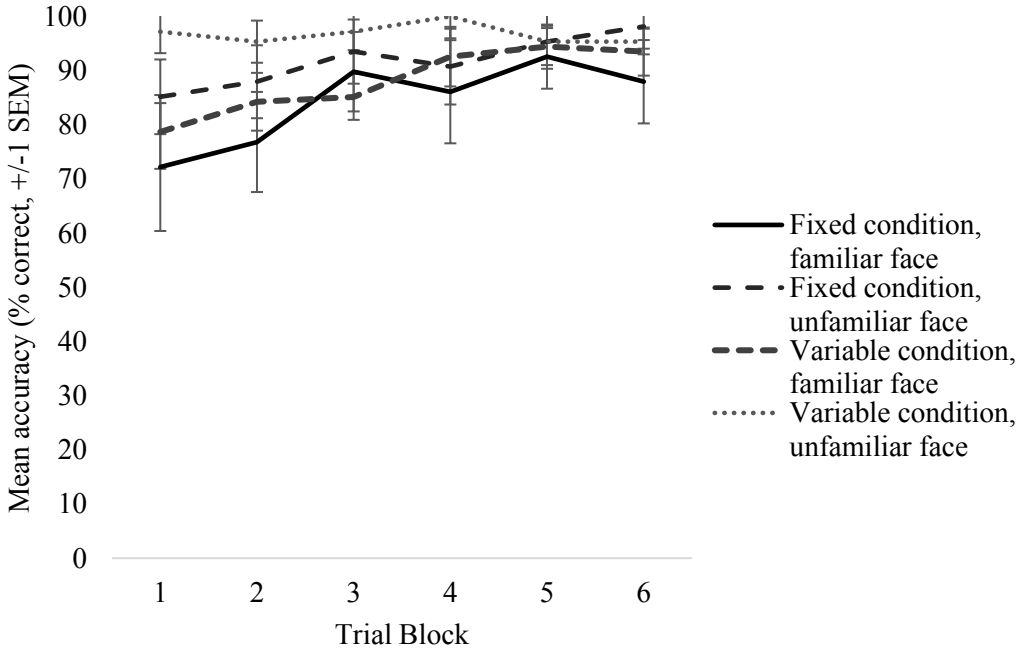


Fig. 35. Mean accuracy in *female* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

As shown in figs. 34 and 35, the results show that there were significant effects of *trial block*, F (2.83, 56.55) = 11.58, $p < .001$, $r = .41$, $\eta^2 = .37$. However, planned contrasts revealed that male participants were only significantly more accurate in the second trial block compared to the first ($p = .03$), and females were only significantly more accurate in the third trial block compared to the second ($p = .01$) No other comparisons were significant.

There were also significant effects of *familiarity*, F (1, 20) = 10.22, $p = .01$, $r = .58$, $\eta^2 = .34$, (familiar: $M = 80.11$, $SE = 3.07$; unfamiliar: $M = 90.76$, $SE = 1.96$) and *participant gender*, F (1, 20) = 5.93, $p = .02$, $r = .48$, (male participant: $M = 80.64$, $SE = 3.41$; female participant: $M = 90.23$, $SE = 1.97$), but no significant effect of *viewing time* on accuracy, F (1, 20) = 1.38, $p = .25$. There was also a three-way interaction between *familiarity*, *trial block* and *viewing time*, F (5, 100) = 2.63, $p = .03$, $\eta^2 = .12$. Finally, as shown in fig. 39, there was also an interaction between *familiarity*, *face gender* and *trial block*, F (5, 100) = 2.28, $p < .05$, $\eta^2 = .10$.

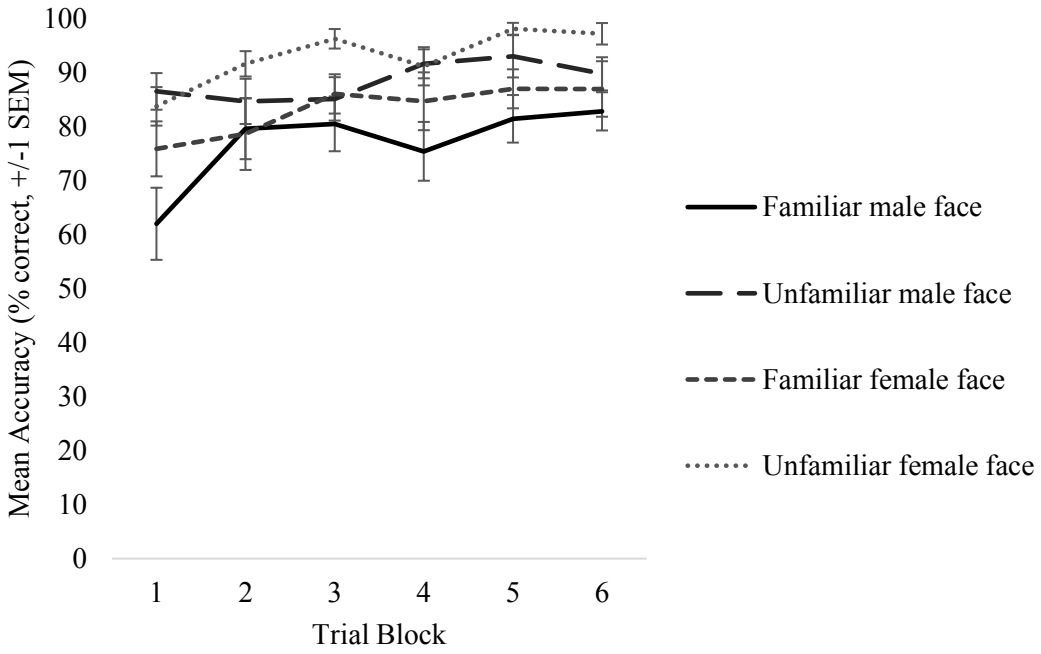


Fig. 36. Mean accuracy in all participants over six sequential trial blocks, as a function of face gender and familiarity.

As in the preceding experiment, the results indicated that participants were better at classifying unfamiliar faces than familiar faces, and that accuracy rates increased significantly across trial blocks. The results also indicate that female participants were more accurate than males. Interestingly, participants were no more accurate in the fixed viewing time condition, although they had longer to look at the images. However, we did find that the patterns of learning faces were different between the two viewing time conditions. In the fixed viewing time condition, participants correctly responded to the *unfamiliar* faces at ceiling levels throughout the experiment, while accuracy in processing the *familiar* faces improved gradually. In contrast, in the variable viewing time condition, accuracy improved gradually with *both* familiar and unfamiliar faces. There were also differences in accuracy as participants responded to familiar and unfamiliar male and female faces: accuracy for unfamiliar male faces did not improve until the fourth trial

block, while the greatest improvements in accuracy were seen for the other face types in the first three trial blocks.

2.9.2. Reaction Times

We did not analyse reaction times, as participants in the fixed viewing time condition could respond whenever they wanted, even after the image had disappeared. In this instance, their response would trigger the drift check.

2.9.3. Pupillary responses

A similar analysis was performed to analyse pupil sizes while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 38.97, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .54$).

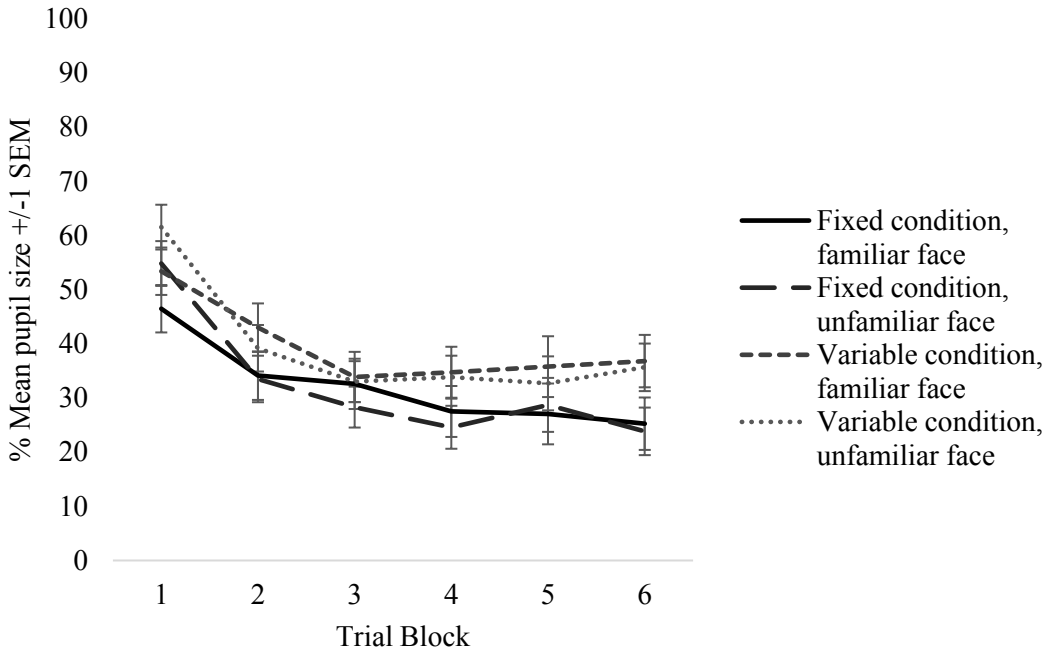


Fig. 37. Mean pupil sizes in *male* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

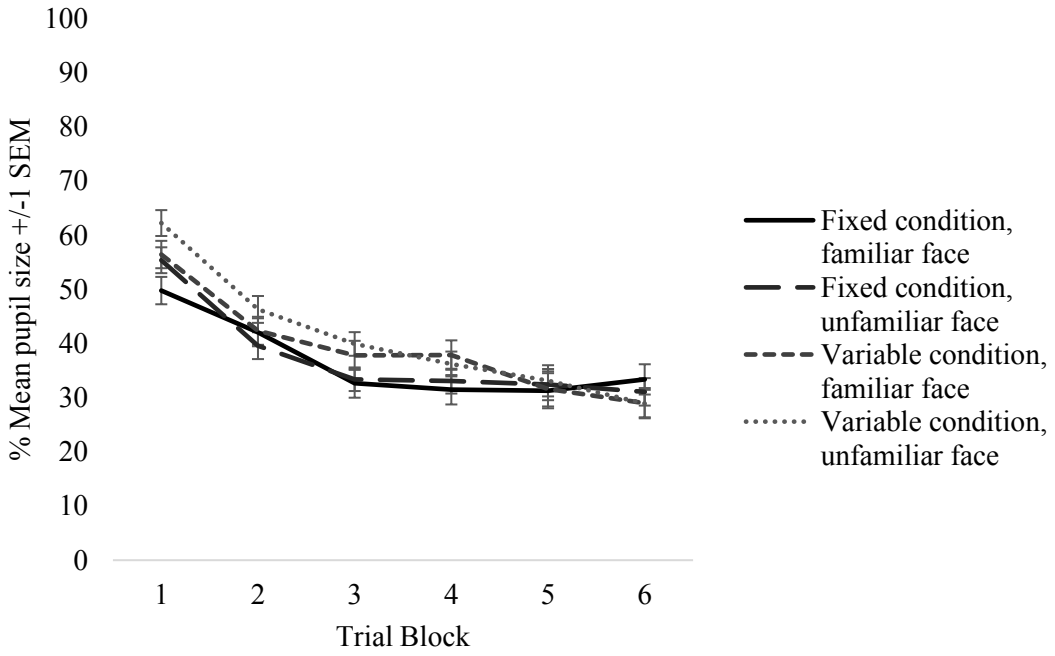


Fig. 38. Mean pupil sizes in *female* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

The results show that there were significant effects of *trial block*, F (2.72, 54.43) = 85.08, $p < .001$, $r = .78$, $\eta^2 = .81$, and an interaction between *trial block* and *familiarity,* F (3.69, 73.85) = 6.29, $p < .001$, $\eta^2 = .24$. Planned contrasts revealed that pupil size was significantly smaller in the second trial block than in the first, and in the third than in the second (both ps < .001). However, there were no other significant differences: in particular, viewing condition was not significant either as a main effect or in interaction with any other variables.

As in both preceding experiments, the results indicated that pupil sizes decreased with each successive trial block. Also, the interaction between *trial block* and *familiarity* indicated that the familiar and unfamiliar faces elicited different pupillary changes over trial blocks, with a steeper initial change for unfamiliar faces compared to familiar faces. However, there were no differences between the viewing time conditions in terms of pupil sizes or pupil size changes over trials, indicating that the viewing time had no effect on pupil sizes.

2.9.4. Fixations

A similar analysis was performed to analyse the number of fixations while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 168.62, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .33$).

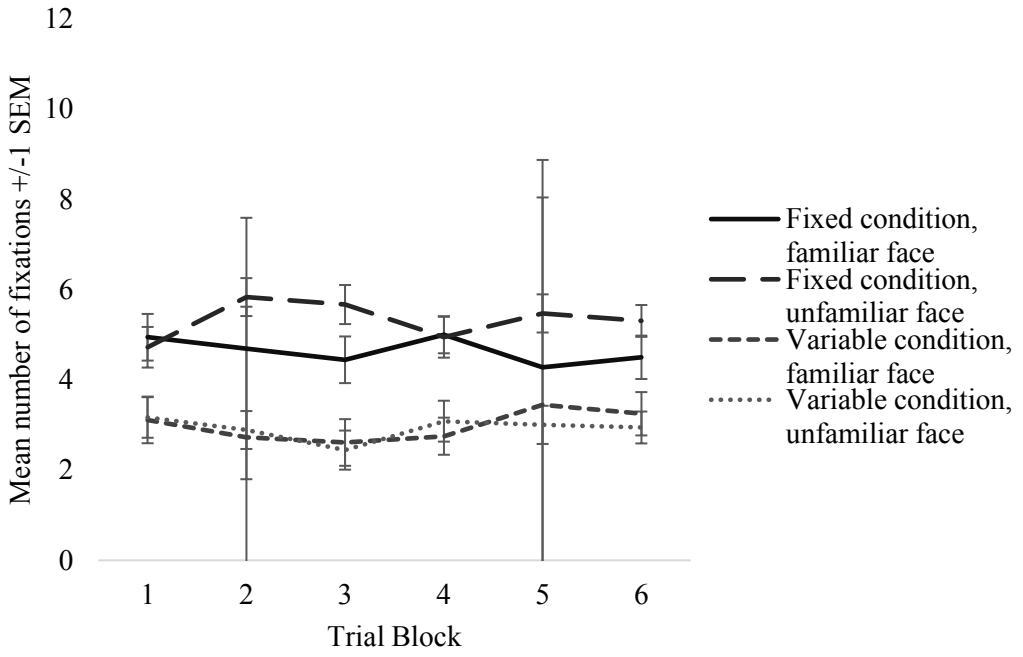


Fig. 39. Mean number of fixations in *male* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

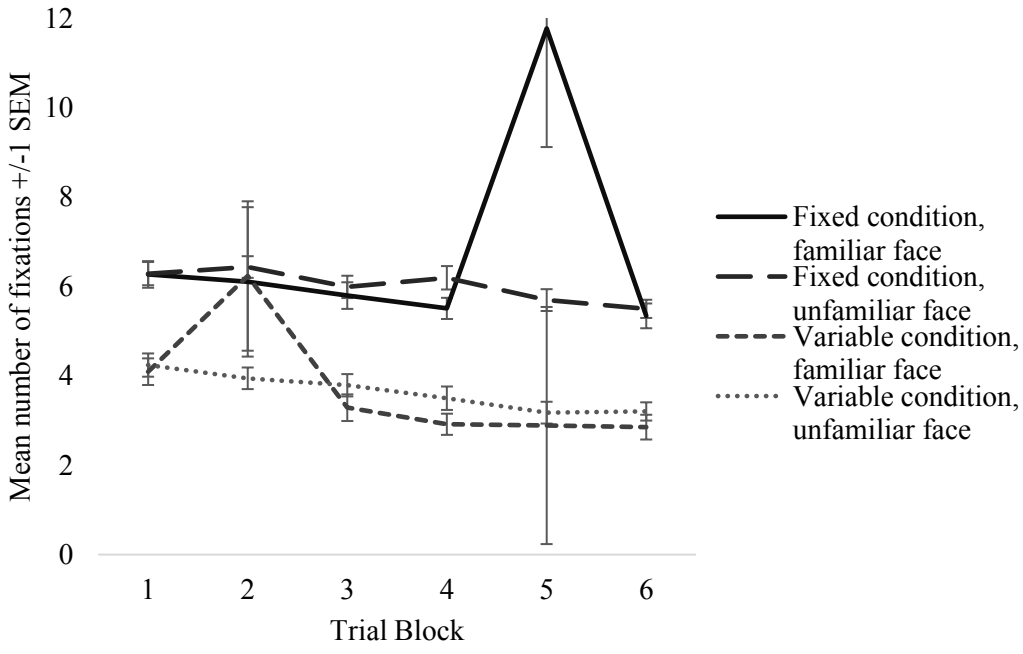


Fig. 40. Mean number of fixations in *female* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

Unlike the preceding experiments, there were no significant effects of *familiarity* or *trial block*, and no interactions. However, there was a significant effect of viewing time, F (1, 20) = 20.57, $p < .001$, $r = .71$: as might be expected, participants in the fixed viewing time condition produced significantly more fixations than in the reaction time condition (fixed: $M = 5.70$, $SE = 0.37$; reaction: $M = 3.31$, $SE = 0.37$).

The results indicate that, all participants made significantly more fixations when they had longer to look at the images, but there was no discernible pattern in the data and the standard errors of the means were very large in some trial blocks. Therefore, it was likely that something unexplained had affected the results. This may due somewhat to the way the data were prepared. The focus of the experiment was to evaluate pupillometry as a measure of face processing (which was compared to the other measures), so we prepared the data accordingly.

The pupillary results showed that there were no outliers. Therefore, we used all participant data for the pupillary analyses *and* the supplementary analyses. An established way to prepare pupillary data with no outliers is to use a mean baseline correction (Mathôt, Fabius, Van Heusden, & Van der Stigchel, 2018). This is the method used in this experiment (the mean baseline correction was converted to zero percent).

Also, using pupillometry, the standard method is to remove trials with blinks, as blinks can distort the results (Kret & Sjak-Shie 2018). Therefore, the pupillary data were first analysed including trials with blinks (and checked for outliers). Then the results were compared those when 'blink' trials had been removed, but there was no difference in the results. The pupil sizes were larger for unfamiliar faces and reduced (in a negative acceleration) during the experiment whether or not trials with blinks were removed. Therefore, all trials were used. This meant that there was not an unnecessary loss of data

(which is a problem for pupillometry), and allowed blink data also to be evaluated in this experiment. Having established that all participants' data could be used including trials with blinks, we did not remove outliers while evaluating the other measures (RTs, accuracy, blinks or fixations). Removing outliers from these other measures would have meant that direct comparisons could not be made between each measure (if different participants were removed for each analysis). This approach only appeared to have affected the results in the present fixation analyses.

2.9.5. Blinks

A similar analysis was performed to analyse the number of blinks while viewing familiar and unfamiliar faces. Mauchly's test indicated that the assumption of sphericity had been violated for *trial block*, χ_2 (14) = 43.28, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .51$).

There were no significant effects. The largest F was for *participant gender*, F (1, 20) = 3.77, $p = .07$, $r = .40$ (females: $M = 0.41$, $SE = 0.08$; males: $M = 0.10$, $SE = 0.14$).

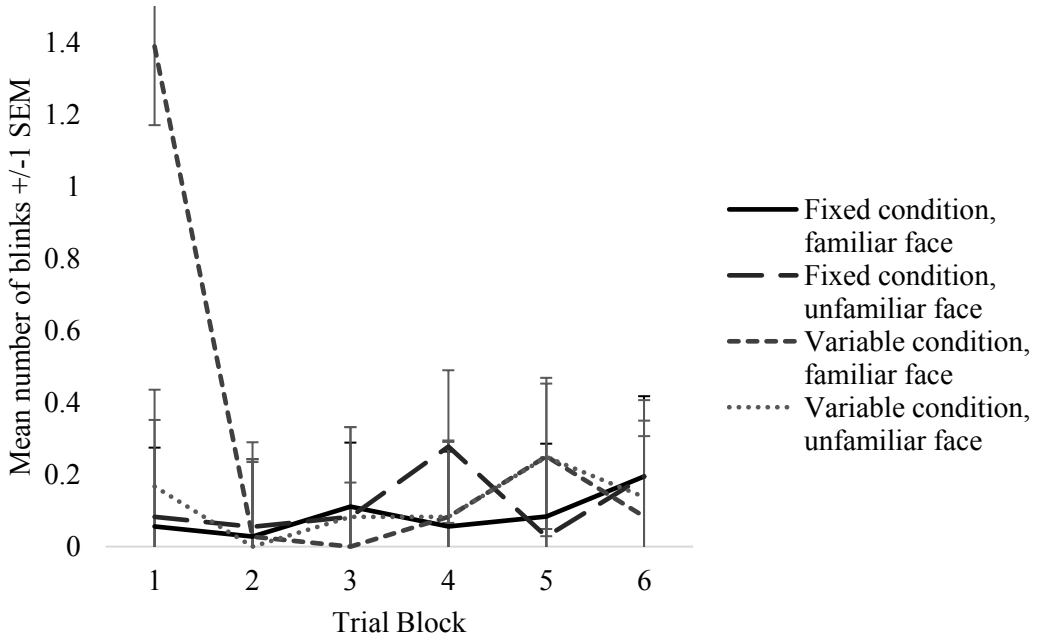


Fig. 41. Mean number of blinks in *male* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

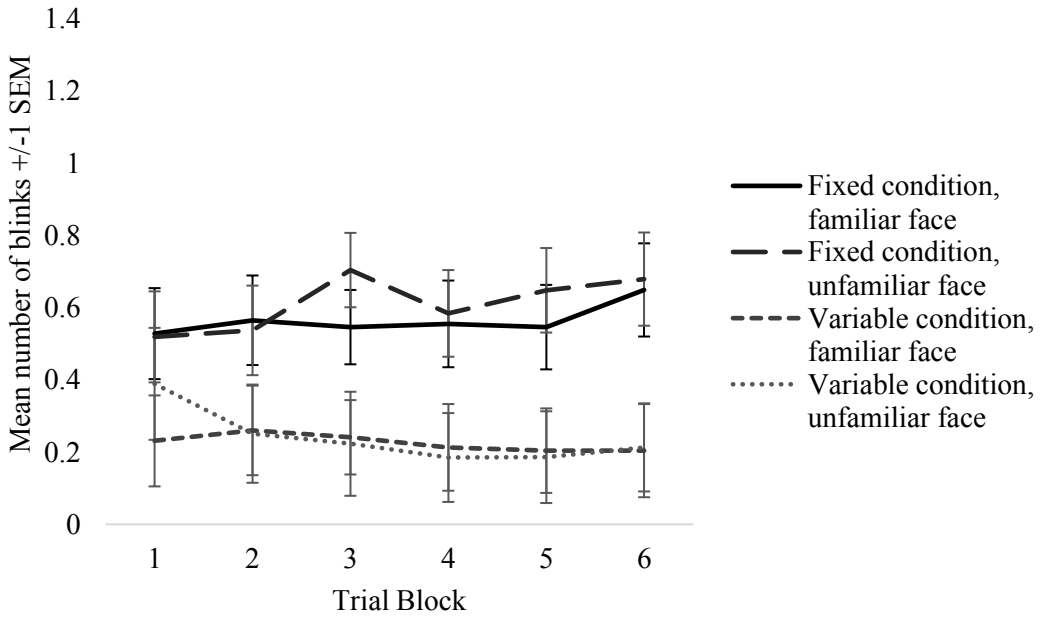


Fig. 42. Mean number of blinks in *female* participants over six sequential trial blocks, as a function of viewing time and face familiarity.

There were no other significant effects. Specifically, *viewing time* had no effect on the number of blinks made by participants, F (1, 20) = 1.29, p = .27.

2.10. Discussion

Experiment 3 investigated whether viewing time might be responsible for some of the effects found in the previous two experiments. Viewing time had no effect on accuracy (participants were equally accurate in both conditions), blinks (they blinked equally often in both conditions) or pupil sizes (pupil sizes were similar in both conditions). This indicated that changes in accuracy, pupil sizes and reaction times all co-occur as learning occurs, all supporting an account of gradual learning. In other words, smaller pupil size changes were not a result of participants' shorter reaction times as familiarity with the faces increased. However, there was a difference between the two viewing conditions for fixations: participants in the fixed viewing time condition made more fixations than those in the variable viewing time condition, indicating that the number of fixations in the preceding two experiments may be an artefact of RTs rather than a measure of cognitive load. The results from Experiment 3 support the idea that pupil size may be a more reliable measure of the cognitive processes associated with familiar and unfamiliar face processing and face learning than the other physiological measures recorded in these experiments. Experiment 3 also clarified some of the gender differences seen in the preceding experiments, as this experiment was not complicated by also testing for effects of age or race. This will be discussed in the following section.

2.11. General Discussion

This exploratory research aimed to investigate whether familiar and unfamiliar face processing were associated with different amounts of mental effort. In two of our experiments (Experiments 2 and 3) accuracy was better when responding to unfamiliar faces (there was no difference in accuracy in Experiment 1), and Experiments 1 and 2 found that participants made more fixations and had bigger pupil sizes when looking at unfamiliar faces. Experiment 1 also found that participants blinked more and were slower when processing unfamiliar faces. Together the RTs and physiological responses seem to suggest that more mental effort was made when classifying the unfamiliar faces as "unfamiliar" than when classifying familiar faces as "familiar". The fact that unfamiliar faces were also generally classified accurately could therefore be attributed to this extra mental effort, but it is more likely that the response bias found for unfamiliar faces affected accuracy (e.g. Bindemann & Sandford, 2011; Jenkins et al., 2011).

This was supported by other data in Experiment 2: participants were less accurate when processing Asian faces, but had larger pupil sizes and made more fixations; whereas they were more accurate when processing male faces, which produced shorter RTs and fewer fixations. This suggests that faces that were easier to classify also required less mental effort. However, Experiment 3 tells a different story again: male participants were less accurate than female participants, and they also made fewer fixations and fewer blinks. It may well be that while familiar and unfamiliar face processing are associated with fluctuations in mental effort, they are also influenced by other cognitive processes, such as engagement in socially-important faces (Keyes & Zalicks, 2016). While we chose experimentally-learnt faces to minimize the salience and valence of faces, it became clear from the comments that participants made during the experiments that some faces ‘jump

out' to some participants more than others (for instance if they remind the participant of someone they know, or if they are more attractive). Thus, while this experiment attempted to find general patterns of face learning, faces cannot be perceived in isolation from the social context.

We also wanted to test whether mental effort could be measured physiologically, and we concluded that it could. As accuracy increased during the experiment, the scores in all our measures decreased. We concluded that the RT and pupillary data were the most successful, as they revealed nuanced effects that were absent in the accuracy data. The fixation and blink data differed very little in absolute terms.

The study also aimed to investigate whether face learning occurs suddenly, suggesting that faces are categorised as either familiar or unfamiliar, or whether they were learned gradually as mental representation became more robust. We found that improvements in accurate face recognition occurred gradually, indicating that the early stages of face learning occur gradually, as reflected in the reduction in RTs and changes in the three physiological responses. Finally, we aimed to see which of the physiological responses provided the most successful measure of cognitive load and whether they could also detect fluctuations in cognitive load associated with the OAB (see Rhodes & Anastasi, 2012 for a review), the ORE (e.g. Michel et al., 2006; Meissner et al., 2013) and the Own Gender Bias (e.g. Wright & Sladden, 2003; Lovén et al., 2011; Lovén et al., 2012). It was found that pupillary responses were the most promising of these physiological measures.

In all three experiments, accuracy improved gradually, and as participants improved in classifying the familiar faces, they also improved at classifying the unfamiliar faces, probably because they used the increasing information they learnt about

the familiar faces to differentiate them from the unfamiliar ones. However, decision responses made for unfamiliar faces were likely to be poor measures of unfamiliar face processing in this experiment, probably due to an 'unfamiliar' response bias (Bindemann & Sandford, 2011; Jenkins et al., 2011). Therefore, the decision responses in this experiment were more likely to be reliable indices of familiar face processing than unfamiliar face processing.

However, there were some interesting findings. For instance, in Experiment 1, female participants were more accurate than males, especially when processing young male faces, providing support for asymmetrical gender biases that were moderated by the OAB (Lovén et al., 2011; Lovén et al., 2012). In Experiment 2, Asian participants improved faster than Caucasian participants and were more accurate overall. They were also considerably better at processing other-race faces than the Caucasian participants were, which supports the idea that both the Own Gender Bias and the ORE can be asymmetrical (Chiroro & Valentine, 1995; Kandel et al., 2016; Meissner & Brigham, 2001).

In Experiments 1 and 2, RTs decreased gradually over successive trial blocks, but an effect of familiarity was only found in Experiment 1. Analysis of the results showed that this was largely due to significantly longer RTs in old participants when processing unfamiliar faces. RTs gradually decreased over time during the first two experiments, indicating that they are fairly reliable in indexing face learning, although they continued to decrease after accuracy improvements had dwindled. Overall, the RT data suggests that they may be more sensitive measures of face learning than decision responses, as they provided more nuanced effects and interactions than the decision responses data. Finally, old participants were slower than young participants and their RTs decreased

more dramatically than those of young participants (old participants' RTs decreased from 2126ms in the first trial block to 1265ms in the sixth, and young participants' RTs decreased from 1002ms in the first trial block to 800ms in the sixth). However, there were no differences in RTs or RT patterns over time between races (Caucasian participants = from 1038ms to 859ms, Asian participants = from 1031ms to 792ms).

Analyses of the results of Experiments 1 and 2 showed partial support for gender differences that require further investigation. In Experiment 1, male participants reacted more slowly than females, particularly when the faces were young; and all participants reacted more slowly to unfamiliar, young female faces than to the other face types. However, in Experiment 2, while participants also reacted more slowly to female faces than male faces, this only occurred when they were *familiar*. Overall, the results suggest that female faces take longer to process than male faces, particularly when combined with other difficult tasks such as other-age face processing. However, as discussed earlier, the female stimuli in Experiment 2 might have been less distinctive than the male stimuli, and this could also have been the case in Experiment 1.

Our focus was on pupillary responses, which appeared to be good indices of familiar and unfamiliar face processing differences in Experiments 1 and 2: pupil sizes were larger when viewing unfamiliar faces. They were also good indirect indices of face learning in all three experiments: pupil sizes reduced in a negative acceleration when looking at familiar and unfamiliar faces, mirroring the trajectory of accuracy well. Additionally, in all three experiments, changes in pupil size over time were steeper initially for unfamiliar faces than familiar faces, suggesting that the mental effort required for classifying unfamiliar faces (compared to familiar faces) was particularly great at the start of the experiment. As discussed earlier, we propose that the task of classifying an

unfamiliar face (based on information built on what was known about the familiar faces) was harder when that information was sparse, and harder than classifying a familiar face about which participants had yet only sparse information.

We also inspected the pupillary responses as measures of own group biases, as research shows that in-group faces are recognised and matched more easily than out-group faces (Meissner & Brigham, 2001). So, we felt that the processing differences might be reflected in the pupillary responses. For instance, if out-group faces were harder to learn or recognise, we would expect pupils to be larger when processing them than in-group faces (see Goldinger & Papesh, 2012 for a review). Alternatively, if in-group faces were more engaging than out-group faces, we would expect pupils to be larger when processing them than out-group faces (e.g. Partala & Surakka, 2003).

We had expected other-age and other-race faces to be associated with greater cognitive load, and to elicit larger pupil sizes as a consequence, but in Experiment 1 there was no interaction between *face age* and *participant age*. Pupil sizes were smaller for young participants than old participants, and young participants had a larger reduction in pupil size than old participants. However, while an account in terms of cognitive load is plausible, the 'larger' pupil sizes of the old participants could have been a physical artefact of autonomic function in older people (Bitsios et al., 1996), as young participants were not significantly more accurate. Therefore, like RTs, pupillary responses do not appear to be good indicators of face learning between participants *of different ages*. This is supported by a lack of differences in pupil sizes between the two races in Experiment 2, although Asian participants were more accurate. We did, however, find partial support for asymmetrical race bias (e.g. Chiroro & Valentine, 1995; Meissner & Brigham, 2001; Kandel et al., 2016): pupil sizes were larger when Caucasian participants looked at Asian

faces compared with Caucasian faces, indicating that processing Asian faces required more mental effort.

There were also no direct effects of gender bias on pupil sizes in either experiment, but there was partial support for a gender bias that warrants further investigation. In Experiment 1, gender biases interacted with age and familiarity: male participants had marginally larger pupil sizes than female participants, and when faces were young and female or familiar and female, they elicited smaller pupil sizes than other face types; and when faces were different combinations of age and gender, pupil sizes reduction had different trajectories. In Experiment 2, there were also different patterns of change in pupillary responses for faces of different genders: pupil sizes reduced in a more linear fashion when looking at female faces.

The first two experiments also found that participants needed to make more fixations when processing unfamiliar than familiar faces, and that they needed fewer fixations as the familiar faces became more familiar, indicating that fixations were also a good indirect index of familiar and unfamiliar face processing and face learning. However, while fixations negatively accelerated in Experiment 1, they reduced linearly in Experiment 2 until the fifth presentation, which was after the point at which improvements in accuracy had dwindled, suggesting that fixations are not ideal indices of real-time face learning.

There was support for an asymmetrical age bias in Experiment 1 (Anastasi & Rhodes, 2005): older participants made more fixations than younger participants, and only the old participants required more fixations to process unfamiliar than familiar faces. Compared to young participants, they also made double the number of fixations on young faces, suggesting that old participants might treat other-age faces as unfamiliar. There

was also support for an asymmetrical race bias in Experiment 2: participants made more fixations when looking at Asian faces compared with Caucasian faces, which may have been an effect of contact (Chiroro & Valentine, 1995), foreign language acquisition (Kandel et al., 2016), because the Caucasian faces were more socially important (Keyes & Zalicks, 2016), or a bias found in Caucasians (Meissner & Brigham, 2001).

Both Experiments 1 and 2 demonstrated indirect and partial support for gender biases (Lovén et al., 2011; Lovén et al., 2012). In Experiment 1, different combinations of familiarity, *face age*, *face gender* and *participant gender* elicited different numbers of fixations, and in Experiment 2, participants made more fixations when looking at female faces. However, in Experiment 3: the number of fixations was higher in the fixed viewing time condition, when participants had longer to look at the images. So, it is plausible that the effects found in Experiments 1 and 2 are artefacts of reaction times rather than direct measures of face processing.

It was only in Experiment 1 that we found that participants also made more blinks when looking at unfamiliar faces: only old participants showed a decrease in the number of blinks over trial blocks. Therefore, the number of blinks made for each image seems to be ineffective at indexing fluctuations in cognitive load associated with learning.

In conclusion, while decision responses appear to be reasonably useful indices of overt familiar face processing, they appear to be unsuccessful measures of unfamiliar face processing in this type of experiment. However, they revealed a partial asymmetrical race bias: Caucasian participants had more difficulty in processing Asian faces than the other way around. This is consistent with previous research on the ORE (e.g. Chiroro & Valentine, 1995; Meissner & Brigham, 2001; Kandel et al., 2016). RTs were probably a subtler behavioural measure of face processing than the decision responses.

Of the three physiological responses, pupillary responses were the most consistent with the accuracy data, as pupil sizes were larger when participants were less accurate, and they were not affected by the RTs. The pupillary data suggest that mental effort was greater when performance was lower, something that was also apparent in some of the interactions. They also indicated that processing unfamiliar faces required more mental effort than familiar faces, and that the extent of this difference diminished quickly at the start of the experiment. However, they were poor measures of the differences in cognitive load between participants, as the pupillary differences between participants from two age groups could be attributed to age-related decline of autonomic function in older people rather than cognitive load, and there were no overall differences when participants were from two different races. In short, pupil sizes provided reasonably reliable measures of the cognitive differences in processing familiar and unfamiliar faces, as they indicated that initial face learning occurs gradually, that some degree of face learning is possible during an experiment, and that face learning is associated with a gradual reduction in cognitive load.

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043–1047.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*, *35*(8), 1089–1105. https://doi.org/10.1068/p5547

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625–627. https://doi.org/10.1068/p7008

Bitsios, P., Prettyman, R., & Szabadi, E. (1996). Changes in autonomic function with age: a study of pupillary kinetics in healthy young and old people. *Age and Ageing*, *25*(6), 432–438.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037//1076-898X.7.3.207

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces: Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Caldara, R., & Abdi, H. (2006). Simulating the 'other-race' effect with autoassociative neural networks: further evidence in favor of the face-space model. *Perception*, *35*(5), 659–670. https://doi.org/10.1068/p5360

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental*

Psychology Section A, *48*(4), 879–894. https://doi.org/10.1080/14640749508401421

CHUK (2009), retrieved Feb 5th 2018, from http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html

Devue, C., & Brédart, S. (2008). Attention to self-referential stimuli: Can I ignore my own face? *Acta Psychologica*, *128*(2), 290–297. https://doi.org/10.1016/j.actpsy.2008.02.004

Devue, C., Van der Stigchel, S., Brédart, S., & Theeuwes, J. (2009). You do not find your own face faster; you just look at it longer. *Cognition*, *111*(1), 114–122. https://doi.org/10.1016/j.cognition.2009.01.003

FEI Face Database (2006), retrieved from http://fei.edu.br/~cet/facedatabase.html, 18 August, 2017.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science, 21*(2), 90-95. https://doi.org/10.1177/0963721412436811

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Harrison, V., & Hole, G. J. (2009). Evidence for a contact-based explanation of the own-age bias in face recognition. *Psychonomic Bulletin & Review*, *16*(2), 264–269. https://doi.org/10.3758/PBR.16.2.264

Hills, P. J., & Pake, J. M. (2013). Eye-tracking the own-race bias in face recognition: Revealing the perceptual and socio-cognitive mechanisms. *Cognition*, *129*(3), 586–597. https://doi.org/10.1016/j.cognition.2013.08.012

Ikehara, C. S., & Crosby, M. E. (2005). Assessing cognitive load with physiological sensors. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (pp. 295a–295a). IEEE.

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kandel, S., Burfin, S., Méary, D., Ruiz-Tada, E., Costa, A., & Pascalis, O. (2016). The impact of early bilingualism on face recognition processes. *Frontiers in Psychology*, *7, 1–9*. https://doi.org/10.3389/fpsyg.2016.01080

Keyes, H., & Zalicks, C. (2016). Socially important faces are processed preferentially to other familiar and unfamiliar faces in a priming task across a range of viewpoints. *PLOS ONE*, *11*(5), e0156350. https://doi.org/10.1371/journal.pone.0156350

Kosaka, H., Omori, M., Iidaka, T., Murata, T., Shimoyama, T., Okada, T., … Wada, Y. (2003). Neural substrates participating in acquisition of facial familiarity: an fMRI study. *NeuroImage*, *20*(3), 1734–1742. https://doi.org/10.1016/S1053-8119(03)00447-6

Kret, M. E., & Sjak-Shie, E. E. (2018). Preprocessing pupil size data: Guidelines and code. *Behavior research methods*, 1-7.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77–100. http://dx.doi.org.ezproxy.sussex.ac.uk/10.1037/0096-1523.34.1.77

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women's own-gender bias in face recognition memory: The role of attention at encoding. *Experimental Psychology*, *58*(4), 333–340. https://doi.org/10.1027/1618-3169/a000100

Lovén, J., Rehnman, J., Wiens, S., Lindholm, T., Peira, N., & Herlitz, A. (2012). Who are you looking at? The influence of face gender on visual attention and memory for own- and other-race faces. *Memory*, *20*(4), 321–331. https://doi.org/10.1080/09658211.2012.658064

Martins, R., & Carvalho, J. M. (2015). Eye blinking as an indicator of fatigue and mental load-a systematic review. *Occupational Safety and Hygiene III*, 231-235.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods*, *50*(1), 94-106.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1473–1483. https://doi.org/10.1080/17470218.2011.575228

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037//1076-8971.7.1.3

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own- and other-race faces. *Visual Cognition*, *21*(9–10), 1287–1305. https://doi.org/10.1080/13506285.2013.832451

Michel, C., Rossion, B., Han, J., Chung, C.-S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608–615.

My Light Meter Pro (2015), retrieved from, https://itunes.apple.com/gb/app/mylightmeter-pro/id583922375?mt=8, 18th August 2017.

The National Eye Institute, 2017), retrieved 25th May, from https://nei.nih.gov/health/dryeye/dryeye

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pilz, K. S., Bülthoff, H. H., & Vuong, Q. C. (2009). Learning influences the encoding of static and dynamic faces and their recognition across different spatial frequencies. *Visual Cognition*, *17*(5), 716–735. https://doi.org/10.1080/13506280802340588

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Proietti, V., Pisacane, A., & Macchi Cassia, V. (2013). Natural experience modulates the processing of older adult faces in young adults and 3-year-old children. *PLoS ONE*, *8*(2), e57499. https://doi.org/10.1371/journal.pone.0057499

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: a PET study of face categorical perception. *Journal of Cognitive Neuroscience*, *13*(7), 1019–1034.

Schluroff, M., Zimmermann, T. E., Freeman Jr, R. B., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary responses to syntactic ambiguity of sentences. *Brain and Language*, *27*(2), 322-344.

Tacikowski, P., & Nowicka, A. (2010). Allocation of attention to self-name and self-face: An ERP study. *Biological Psychology*, *84*(2), 318–324. https://doi.org/10.1016/j.biopsycho.2010.03.009

Tong, F., & Nakayama, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1016–1035. http://dx.doi.org/10.1037/0096-1523.25.4.1016

Van Belle, G. (2010). Fixation patterns during recognition of personally familiar and unfamiliar faces. *Frontiers in Psychology*. *1*, 20. https://doi.org/10.3389/fpsyg.2010.00020

VidTIMIT (2009) C. Sanderson and B.C. Lovell Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. Lecture Notes in Computer Science (LNCS), Vol. 5558, pp. 199-208, 2009. Retrieved from, , 18th August 2017.

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica*, *114*(1), 101–114. https://doi.org/10.1016/S0001-6918(03)00052-0

Zimmermann, F. G. S., & Eimer, M. (2013). Face learning and the emergence of view-independent face recognition: An event-related brain potential study. *Neuropsychologia*, *51*(7), 1320–1329. https://doi.org/10.1016/j.neuropsychologia.2013.03.028

CHAPTER 3. ENGAGING IN SELF-INTEREST: MEASURING OWN FACE PROCESSING WITH PUPILLOMETRY.

Abstract

Pupil sizes change as people look at different types of faces. This study investigated whether pupillary changes occurred as participants were presented with unfamiliar faces, personally-familiar faces, and their own face, and asked what gave rise to them. We evaluated two theories that have been measured using pupillometry. Cognitive load theory suggests that it is less effortful to process familiar faces than unfamiliar faces (resulting in smaller pupil sizes in response to familiar faces). Cognitive engagement theory suggests that familiar faces are more engaging than unfamiliar faces (eliciting larger pupil sizes in response to familiar faces). We found that pupil sizes were significantly larger when viewing own faces compared to the other face types, suggesting that cognitive engagement offers a plausible account of face processing, which was supported by fixation data. However, the results were also consistent with an interpretation in terms of memory strength theory, which indicates that pupil sizes get larger as memory gets stronger, supporting previous research.

Studies show that pupil size is affected by factors other than simply luminance (Binda, Pereverzeva, & Murray, 2014). Cognitive load loosely refers to the amount of mental effort required to perform a task (see Ayres & Paas, 2012; Murphy, Groeger, & Greene, 2016, for a review), and it appears that pupillary responses are associated with cognitive load, as pupil sizes are larger when cognitive load is high and smaller when cognitive load is low (Jainta & Baccino, 2010; Piquado, Isaacowitz, & Wingfield, 2010; Chen & Epps, 2014; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014).

Cognitive load theory (Ayres & Paas, 2012; Murphy et al., 2016) applies well to face processing, as face recognition research has long shown that recognising faces that have only been seen briefly before requires more mental effort and is less successful than recognising faces that are highly familiar (see Hancock, Bruce, & Burton, 2000, for a review). For example, in a face matching study using CCTV images and comparison photographs, Bruce, Henderson, Newman, and Burton (2001) found that accuracy was at approximately 75% when the stimuli were unfamiliar faces, but this rose to approximately 90% when the stimuli were familiar faces. Indeed, it has also been found that pupil sizes are larger when looking at other-race faces (that are harder to recognise and learn) than own race-faces (Goldinger, He, & Papesh, 2009; Wu, Laeng, & Magnussen, 2012).

Cognitive engagement is based upon the premise that socially-important objects or those containing emotional content will be more engaging than objects with no emotional content or social-importance. It has been established that cognitive engagement affects pupil size, as engaging stimuli result in larger pupils (Laeng & Falkenberg, 2007) and unengaging stimuli result in smaller pupils (Bradley, Miccoli, Escrig, & Lang, 2008). This is supported by other research, for example, when presented

with attractive or emotional stimuli, pupils are larger than with neutral stimuli (e.g. Partala & Surakka, 2003; Laeng & Falkenberg, 2007; Bradley et al.; Võ et al., 2008; Prehn, Heekeren, & Van der Meer, 2011; Snowden et al., 2016), and large pupils are associated with goal-seeking and decisions that result in reward (Satterthwaite et al., 2007; Mathôt, Siebold, Donk, & Vitu, 2015). These studies combine to suggest that large pupil sizes are associated with cognitively engaging stimuli. Therefore, it appears overall that while cognitive load does explain some changes in pupil size, cognitive engagement also accounts for pupillary responses.

The studies described above also suggest that an account of pupillary change in terms of cognitive engagement could be applied to face processing, as different faces have different levels of social importance for the observer (Keyes & Zalicks, 2016), that are unequally loaded with emotional, motivational, contextual and social content. Thus, it is likely that the degree to which a face engages a person will affect pupil sizes. This could override changes that occur as a consequence of cognitive load.

However, cognitive load and cognitive engagement need to be teased apart to understand the complex issue of face recognition. Berggren, Koster, and Derakshan (2012) suggest that the constructs can be separated, as they found that cognitive load does not affect emotion processing. Buetti and Lleras (2016) suggest that cognitive engagement can affect distractibility more than cognitive load. They gave participants maths tasks of different degrees of difficulty while distracting them with images. They found that different degrees of cognitive load were not associated with how much participants were distracted by the images. However, the degree of engagement that participants had with the task did: participants who were engaged with the task were less

distracted than those who were not. This indicates that cognitive engagement can over-ride cognitive load when it comes to being distracted.

Face recognition research indicates that familiar and unfamiliar face processing are different in some respects (Ellis, Shepherd, & Davies, 1979; see Johnston & Edmonds, 2009 for a review). Much research demonstrates that recognising or even matching unfamiliar faces is much more difficult than doing so with familiar faces (Ellis et al. 1979; Bruce et al., 2001), and that there are even more difficulties when the unfamiliar faces differ in "race" from the person processing them (Meissner & Brigham, 2001), age (Anastasi & Rhodes, 2005), or when different lighting or viewpoints are involved (Burton, Jenkins, & Schweinberger, 2011). This may be due to differences in processing (discussed in Collishaw & Hole, 2000), due to there being less information available about the unfamiliar faces that can be used to recognise them (Longmore, Liu, & Young, 2008). Some research has investigated face learning (how an unfamiliar face becomes familiar), and much of this suggests that face learning is gradual (Kosaka et al., 2003). In other words, faces become more robustly represented over time and exposure (Burton, Jenkins, Hancock, & White, 2005). Also, faces seen from multiple viewpoints and in different lighting are more easily learnt than those only seen from limited viewpoints (Longmore et al., 2008). Therefore, the faces of well-known people are easily recognised in different lighting and from different viewpoints, but faces only seen briefly and from limited viewpoints are less likely to be recognised when seen again from another viewpoint.

All this suggests that familiar faces fall on a spectrum of familiarity, and that those that are most familiar should be recognised more easily than those on the other end of the spectrum. However, these accounts fail to explore the role of cognitive engagement, in

other words, whether faces that are more familiar engage people more or less than unfamiliar faces. Previous research suggests that familiar faces would be more engaging, as they contain more socially-important information than unfamiliar faces (Keyes & Zalicks, 2016). However, in a situation where a face appears unexpectedly among other faces, it might be more cognitively engaging than the expected faces, regardless of familiarity (Meyer, Reisenzein, & Schützwohl, 1997; Lorini & Castelfranchi, 2007, for reviews on surprise). Therefore, when people look at faces, there are probably more fluctuations in cognitive engagement that are unrelated to familiarity than there are in cognitive load. This is because socially-important information is overlapping, complex and nuanced (e.g. fear, attraction, emotion, motivation, reward), while degrees of familiarity from novel faces to highly familiar ones should be fairly linear.

Previous research has usually investigated familiar face processing using personally-familiar faces, famous faces or experimentally-learnt faces. Unfamiliar faces are usually entirely novel, or have only briefly been presented in an experiment. The issues with famous faces include that they are generally only experienced two-dimensionally, images of them can be 'iconic', which might test pictorial recognition rather than face recognition (Carbon, 2008; & Burton, 2013), and images of actors can be much more varied than those of e.g. politicians, as actors often change their appearance for different roles. The issue with experimentally-learnt faces is that while some learning can occur during an experiment (Pilz, Bülthoff, & Vuong, 2009), they cannot be as robustly represented as highly familiar faces (Tong & Nakayama, 1999). Personally-familiar faces are probably the most robustly represented of those described above, but can be problematic in experiments, as faces that are highly familiar for one participant may be unfamiliar to another participant, and they are associated with personal memories and emotions (Keyes & Zalicks, 2016). Finally, unfamiliar faces can be difficult to test,

as participants cannot ‘recognise’ a face that has never been seen before. However, they can be categorised as unfamiliar (Rossion, Schiltz, Robaye, Pirenne, & Crommelinck, 2001), and are useful in face-matching tasks, when two different images are paired, and the participant has to decide whether they are both images of the same person, or if they are images of two different people (Jenkins, White, Van Montfort, & Burton, 2011).

Own face recognition has also been investigated, but relatively rarely. Research suggests that people have a bias for their own face over other face types. For example, Devue, Van der Stigchel, Brédart, and Theeuwes (2009) found that people looked longer at their own face, and Ninomiya, Onitsuka, Chen, Sato, and Tashiro (1998) found a larger P300 response to own faces than to unfamiliar faces. Kircher et al. (2001) found that (compared to unfamiliar faces) there was increased activity in the right limbic region, left prefrontal cortex, and superior frontal cortex when participants saw their own face, whereas viewing images of their partner produced increased activity only in the right insula. Another study found increased brain activation in the right anterior insula and left inferior parietal lobe, but less activation in the right posterior cingulate/precuneus when own faces were processed, compared with other familiar faces (Ramasubbu et al., 2011).

Previous research suggests that own faces are more difficult to process than other familiar faces, as people generally only see themselves in the mirror. So, the assumption is either that photographs that are not mirror-reversed (veridical) would appear unfamiliar and need to be mentally ‘flipped’, or that the mirror-reversal is merely an uncharacteristic view of their face that involves greater effort (cognitive load) to process in itself. Indeed, it appears that own faces are associated with increases in the right hemisphere (the inferior frontal gyrus and inferior parietal lobule), when presented veridically, that are matched

to stored representations of the self, while other faces are involved in midline brain structures (Uddin, Kaplan, Molnar-Szakacs, Zaidel, & Iacoboni, 2005).

However, Tong and Nakayama (1999) used own faces as examples of robustly-represented face stimuli that were compared to novel and less robustly-represented experimentally-learnt faces, and found that own faces are more robustly represented than newly familiarised faces. Also, more recent research suggests that people are increasingly used to seeing their own images (static and moving) in the mirror, in selfies (photographs of themselves on their mobile phones) (Murray, 2015), and on social media, so they should nowadays be more familiar than other familiar face types, making veridical own face images the easiest to process. For instance, it has been found that own faces are processed faster than unfamiliar faces, even when the unfamiliar faces have been presented multiple times, when the own face is presented from profile or inverted views, or is presented with or without hair (Tong & Nakayama, 1999). It has also been shown that own faces also produce less of a face after-effect than other faces, suggesting own faces are particularly robustly represented (Laurence & Hole, 2011).

By presenting participants with unfamiliar, personally-familiar, and own faces, we will investigate how well changes in pupil size can differentiate between face types. We also hope to understand more about the cognitive process of recognising different face types, particularly own faces; and to tease apart cognitive engagement and cognitive load accounts of pupil size changes. The study used psychology faculty members from two universities as stimuli, with psychology faculty members from both universities as participants, so the personally-familiar faces were well-known colleagues. The study was conducted by two researchers, one from Kent and one from Sussex. All participants used social media for social and work-related activities. They also generally received instant

notifications from their smart device when they were mentioned in a post, invited to an event, and were 'tagged' etc. "Tagging' is when a friend on social media posts an image of you and tags your name to that image, so that it appears automatically in your account, and those of your other social media friends (depending on account settings).

Therefore, we predict that if pupil size changes are caused by variations in cognitive load, pupil sizes will be largest for unfamiliar faces, where the cognitive load is greatest, medium for the personally-familiar faces, and smallest for own faces, where the mental representation of the face should be the most robust. However, if pupil size changes are caused by variations in cognitive engagement, they should be largest when participants view their own face and smallest for the unfamiliar faces.

As mentioned above, cognitive engagement is probably also affected by surprise or anticipation (see Meyer et al., 1997; Lorini & Castelfranchi, 2007 for reviews). In other words, a person would probably be more engaged with a particular face if it appeared unexpectedly, among expected faces. Conversely, if a person was expecting to see a face, it might be particularly engaging when it appeared (Proulx, Sleegers, & Tritt, 2017). So, we tested this by dividing participants into three conditions. First, we tested participants who had previously given consent for their photograph to be used in an unspecified face recognition experiment. Therefore, participants in this condition generally *anticipated* seeing their own face in the experiment. We called this condition "Consented". In another condition, participants who had previously given their consent for us to use their photograph were told explicitly before starting the experiment that this photograph would be included in the experiment. Therefore, these participants *knew* that they would see their own face. We called this condition "Aware". The final condition was labelled "Unaware". These participants came to the experiment without having given consent for

us to use their photograph, but knew that the experiment contained images of faculty members. Therefore, these participants appeared to be *surprised* to see their own face in the experiment. We predicted that the pupil sizes when viewing the own face would be largest for Unaware participants and Aware participants compared to Consented participants, but made no prediction between the Aware and Unaware conditions.

3.1. Method

3.2.1. Participants

Fifty-one participants with normal or corrected to normal vision were recruited from the psychology faculties at the University of Sussex and at the University of Kent. Nine were subsequently excluded due to technical problems. This left forty-two participants (12 males and 30 females) aged between twenty-two and forty-three (M = 28.10, SD = 5.31). Sussex participants included 6 males and 15 females, aged between twenty-thee and forty-three (M = 28.38, SD = 5.90), and Kent participants included 6 males and 15 females, aged between twenty-two and forty (M = 27.81, SD = 4.78).

Participant were grouped into one of three conditions: “Consented”; there were 14 participants in this condition, 6 males and 8 females, aged between 24 and 40 (M = 28.93, SD = 5.31); “Aware”; there were 15 participants in this condition, 4 males and 11 females, aged between 22 and 43 (M = 27.60, SD = 6.17); and “Unaware”; there were 13 participants in this condition, 2 males and 11 females, aged between 23 and 40 (M = 27.77, SD = 4.49).

3.2.2. Apparatus and Materials

The stimuli consisted of twenty-four images of University of Sussex psychology faculty members, and twenty-four images of University of Kent psychology faculty members. Faces that were taken from the same university as the participant were designated as "familiar", faces from the other university were designated as "unfamiliar", and each participant also saw an image of their own face. Own face images were presented veridically for all participants (i.e. they were not mirror-reversed). Therefore, when participants viewed their own face, they were viewing it left-right reversed compared with the way they would see it in the mirror. All images were taken from university web profile pages or Facebook profiles and were cropped and matched for size (13.3cm x 17.5cm) and resolution (390 x 503 pixels). Image luminosity was not altered, but room luminosity was controlled by drawing the room's blinds. The presentation order of the images was randomised for each participant. They were displayed at an approximate distance of 60cm from the chin rest (although this was adjusted for each participant).

Experiment Builder was run on a desktop EyeLink 1000 eye-tracker that uses infrared illumination to record pupil data, and a 21.5 inch iMac computer. The eye-tracker can accommodate small head movements, pupillary hippus (an abnormal rhythmic spasm of the iris (McGraw-Hill, 2002)), pupillary wobble and blinks. Further head movements were stabilised by using a chin rest. The right eye was tracked for all participants.

3.2.3. Design

This study used a mixed design: independent measures on *university*, with two levels (Kent and Sussex), and *condition*, with three levels (consented, aware, and unaware); and repeated measures on *face type* (with three levels: unfamiliar, familiar and own face). There were two dependent variables for behavioural responses: accuracy and

reaction times; and three for physiological responses: pupil size (calculated as percentages of each participant's overall pupil size range during the experiment), fixation counts and blink counts.

3.2.4. Procedure

The participants were briefed before placing their chins on the chin rest. Their eye movements were calibrated to nine points on the screen. They were then asked to press any key to display an instruction page. This was followed by a drift check, designed to monitor gaze throughout the experiment. The drift check was repeated between successive trials, and involved looking at a black dot on a white screen. The first drift check was followed by a practice session, which contained four faces, two from each university. Participants were asked to determine as quickly as possible whether each face was from Kent or Sussex university by clicking "K" for Kent and "S" for Sussex.

After completing the practice session, if the participants felt that they understood the task, they completed the test stage. This contained images of twenty-four individuals from each university. Again, participants were asked to determine as quickly as possible whether the face was from Kent or Sussex university by pressing either "K" or "S". The eye-tracker recorded eye movements, pupil sizes, blinks, fixations, RTs (in milliseconds), and key-press responses as participants viewed the images.

3.2. Results

The first two analyses investigated behavioural measures of face processing: response accuracy and RTs.

3.2.1. Accuracy

We examined participants' accuracy in deciding whether a face was familiar or unfamiliar to see how participants for each university performed for both face types, taking into account whether they were aware that their own face would be shown to them or not. However, for this first analysis on accuracy, we excluded own face data, as all participants responded correctly to their own face. So, we performed a three-way ANOVA with repeated measures on *face type* (with two levels: unfamiliar and familiar), and independent measures on *university*, (two levels: Kent and Sussex), and *condition* (three levels: consented, aware, and unaware).

The results show that there was a significant main effect of accuracy on *face type*, F (1, 36) = 10.47, p = .01, r = .47: participants were significantly more accurate in responding to *unfamiliar* faces. There was also a significant interaction between *face type* and *condition*, F (2, 36) = 6.84, p = .01, η^2 = .25, and a three-way interaction between *face type*, *condition* and *university*, F (2, 36) = 4.77, p = .02, η^2 = .19.

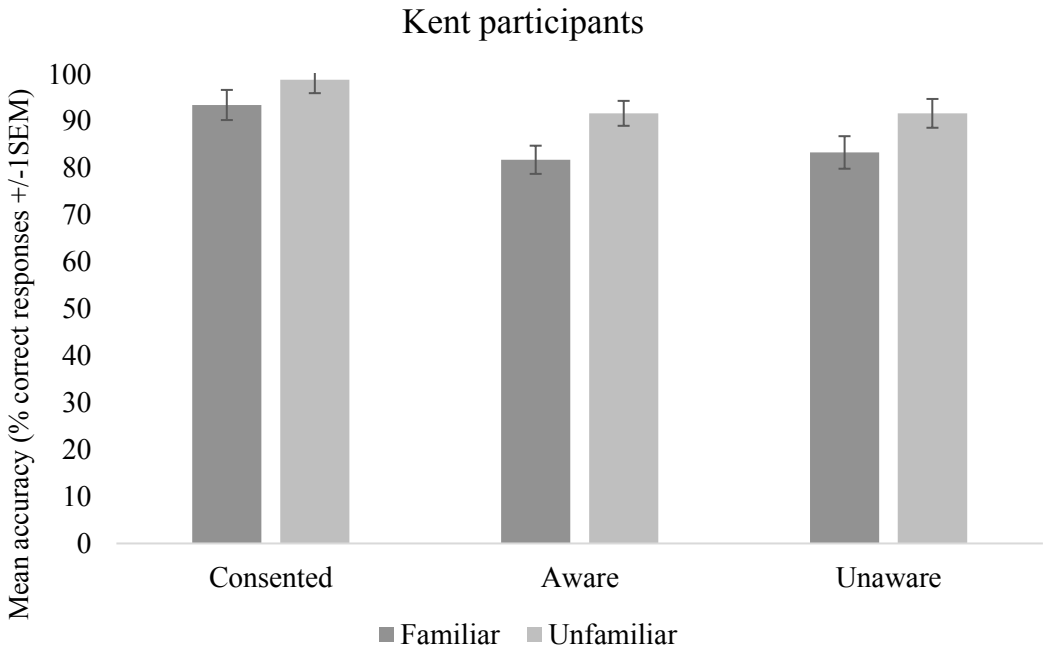


Fig. 43. Response accuracy while viewing familiar and unfamiliar faces, as a function of condition, in Kent participants.

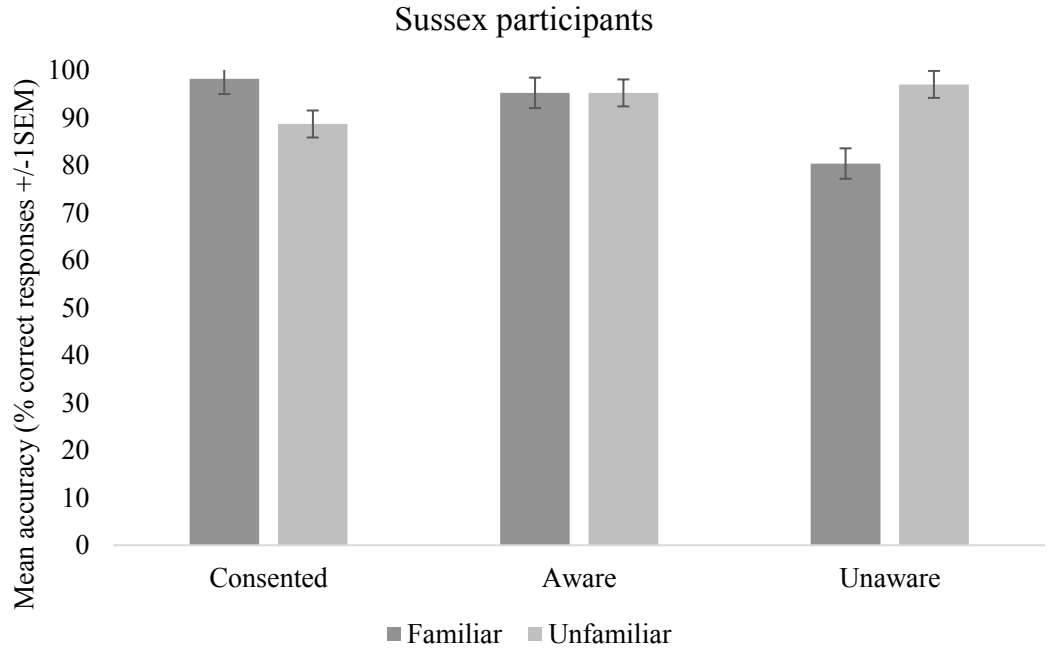


Fig. 44. Response accuracy while viewing familiar and unfamiliar faces, as a function of condition, in Sussex participants.

Therefore, we conducted *t*-tests to see where the differences were. In ***Kent*** participants those in the *consented* and *aware* conditions were significantly more accurate when responding to unfamiliar faces than familiar faces: *consented*, $t(6) = 3.24$, $p = .02$, $r = .59$ (familiar: $M = 88.69$, $SE = 2.36$, unfamiliar: $M = 97.02$, $SE = 0.77$); *aware*, $t(7) = 3.21$, $p = .02$, $r = .56$ (familiar: $M = 78.29$, $SE = 4.39$, unfamiliar: $M = 91.25$, $SE = 3.37$). However, there was no significant difference between *face types* in the *unaware* condition, $t(5) = 2.10$, $p = .09$ (familiar: $M = 81.67$, $SE = 4.59$, unfamiliar: $M = 90.83$, $SE = 4.36$).

However, in ***Sussex*** participants there was a significant difference between *face types* in the *unaware* condition, $t(7) = 3.73$, $p = .01$, $r = .35$ (familiar: $M = 77.50$, $SE = 5.09$, unfamiliar: $M = 96.74$, $SE = 1.70$). However, there were no significant differences between *face types* in the *consented* and *aware* conditions: *consented*, $t(6) = 2.00$, $p = .09$ (familiar: $M = 89.28$, $SE = 3.73$, unfamiliar: $M = 97.62$, $SE = 1.24$); *aware*, $t(6) = 0.03$, $p = .98$ (familiar: $M = 94.88$, $SE = 1.97$, unfamiliar: $M = 94.76$, $SE = 2.73$).

3.3. Reaction times (RTs)

To examine participants' reaction times as they decided whether a face was familiar or unfamiliar, a three-way mixed ANOVA was performed. This had independent measures on *university*, with two levels (Kent and Sussex), and *condition*, with three levels (consented, aware, and unaware); and repeated measures on *face type* (with three levels: unfamiliar, familiar and own face). Mauchly's test indicated that the assumption of sphericity had been violated. Therefore, Greenhouse-Geisser corrected tests are reported ($\varepsilon = .52$): $\chi_2(2) = 93.03$, $p < .001$. There was no significant main effect of *face type*, F (1.04, 37.31) = 0.54, p = .48. There were also no other significant effects. However, the variation in the own-face condition was markedly larger than in the other conditions.

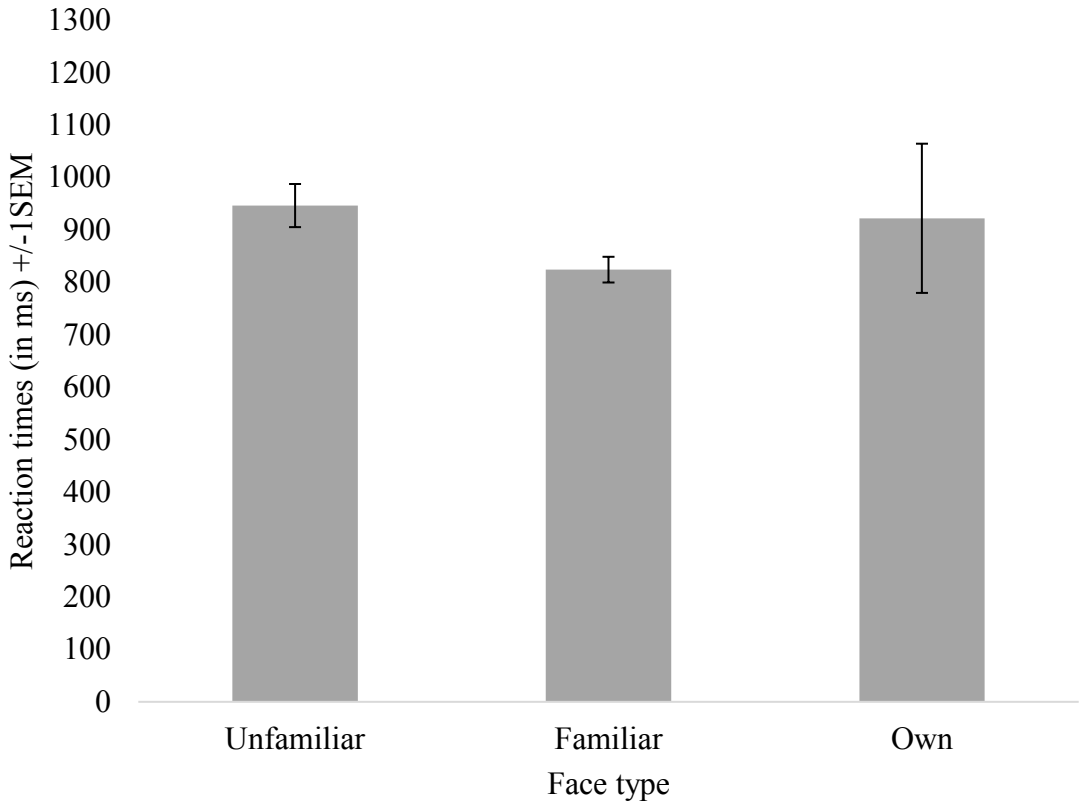


Fig. 45. Reaction times while viewing own, familiar and unfamiliar faces.

The following analyses investigate physiological responses to face processing: pupil sizes, fixations, and blinks.

3.4. Pupillary responses

3.4.1. Pupil sizes

In order to standardise pupil sizes between participants, a mean pupil size for each face was obtained from the eye-tracker. This was converted to a percentage of the pupil size change observed for each participant during the experiment. Percentages were calculated separately for each participant by identifying the face with the largest mean pupil size and the face with the smallest mean pupil size, and calculating the difference

between them. The mean pupil size for each face was then calculated as a percentage of that difference.

Three pupil size measurements were taken from each participant: the mean pupil size for unfamiliar faces; the mean pupil size for familiar faces; and the (single) pupil size for the participant's own face.

To examine participants' pupil sizes as they decided whether a face was familiar or unfamiliar, a three-way mixed ANOVA was performed. This had independent measures on *university*, with two levels (Kent and Sussex), and *condition*, with three levels (consented, aware, and unaware); and repeated measures on *face type* (with three levels: unfamiliar, familiar and own face). Mauchly's test indicated that the assumption of sphericity had been violated for *face type*. Therefore, Greenhouse-Geisser corrected tests are reported ($\varepsilon = .55$): $\chi_2(2) = 59.30$, $p < .001$. There was a significant main effect of *face type* on pupil size, $F(1.11, 40.94) = 4.47$, $p = .04$, $r = .31$, $\eta^2 = .18$ (unfamiliar: $M = 38.53$, $SE = 1.39$, familiar: $M = 36.76$, $SE = 1.53$, own: $M = 41.85$, $SE = 2.74$), but no other significant results.

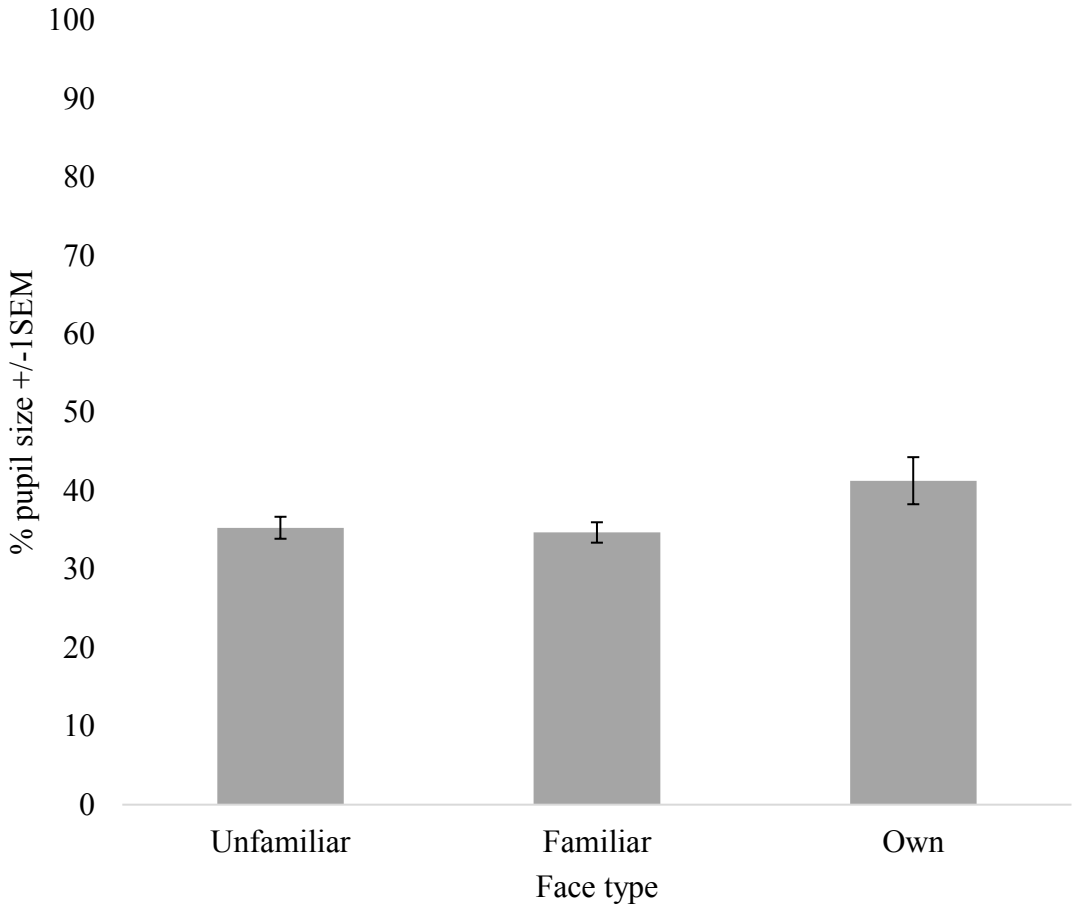


Fig. 46. Pupillary changes while viewing own, familiar and unfamiliar faces.

3.5. Fixations

A similar analysis was performed on the number of fixations produced as participants viewed the three face types. Mauchly's test indicated that the assumption of sphericity had been violated. Therefore, Greenhouse-Geisser corrected tests are reported ($\varepsilon = .73$): $\chi_2 (2) = 16.55$, $p < .001$. There was a significant main effect of *face type*, $F (1.45, 52.29) = 19.36$, $p < .001$, $r = .46$, $\eta^2 = .34$, but no other significant main effects or interactions.

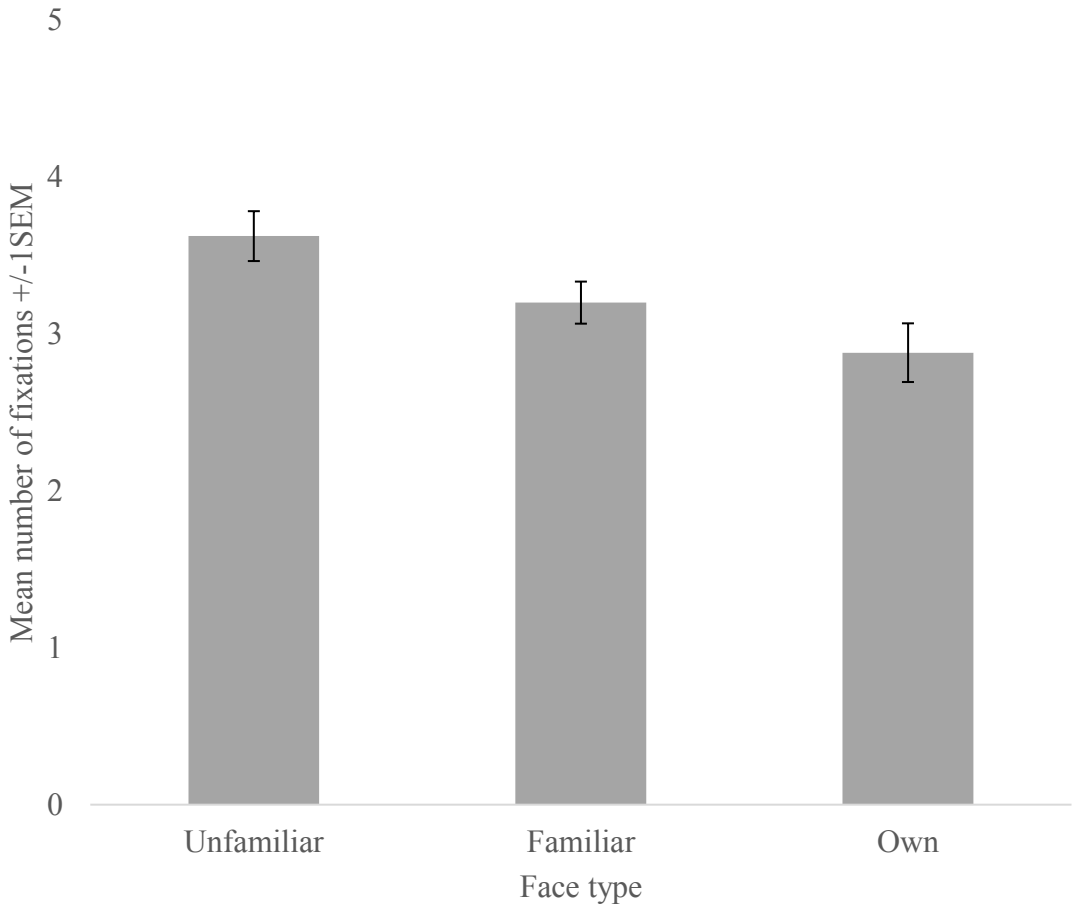


Fig. 47. Number of fixations while viewing own, familiar and unfamiliar faces.

As can be seen from fig. 47, participants made significantly fewer fixations as faces became more familiar.

3.6. Blinks

A similar analysis was performed on the number of blinks produced as participants from each university viewed the three face types. Mauchly's test indicated that the assumption of sphericity had been violated. Therefore, Greenhouse-Geisser corrected tests are reported ($\varepsilon = .57$): $\chi_2 (2) = 51.58$, $p < .001$. There was no significant

effect of *face type* on the number of blinks, F (1.13,40.66) = 2.96, p = .09. There were also no significant interactions.

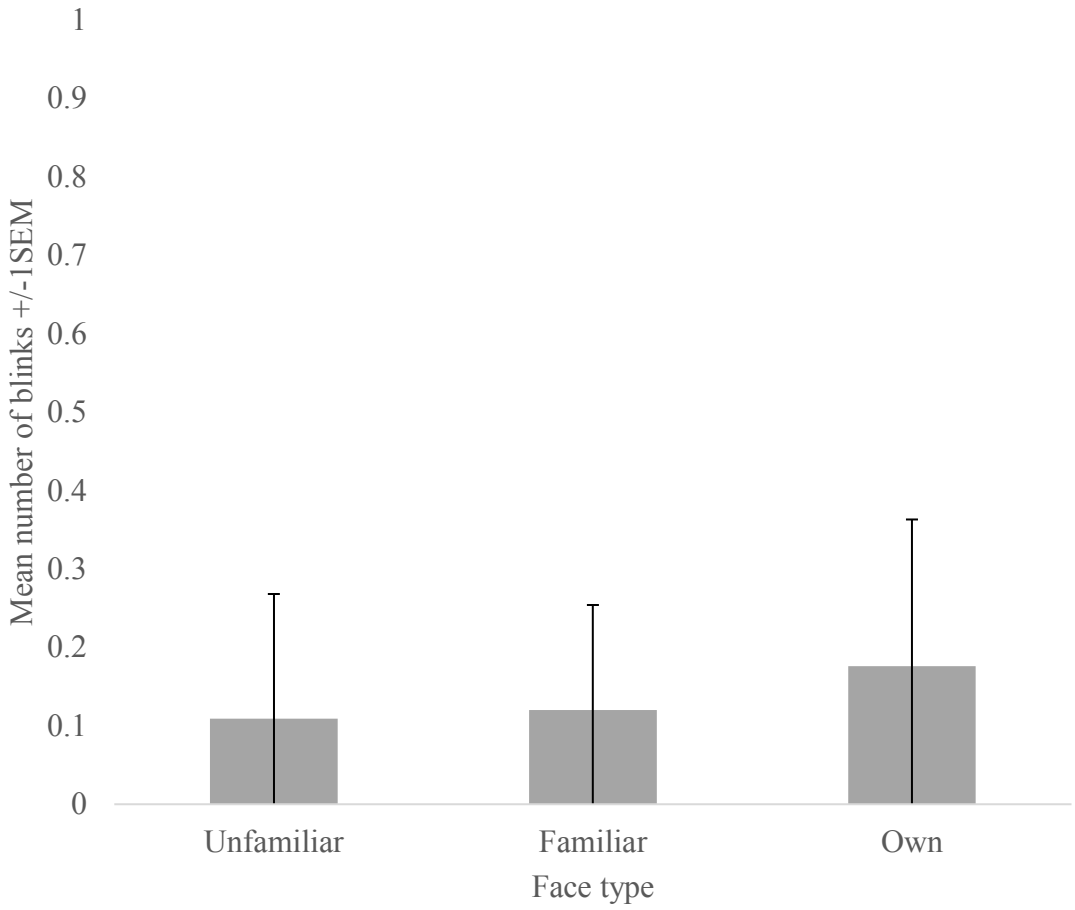


Fig. 48. Number of blinks while viewing own, familiar and unfamiliar faces.

3.7. Discussion

We investigated whether pupil sizes were different when viewing familiar, unfamiliar and own faces. Our primary aim was to see whether pupil sizes changed when looking at different face types, and to evaluate them as reliable measures of face processing. We compared them to our findings of more traditional behavioural measures of face processing (accuracy and RTs), and against other physiological measures

(fixations and blinks); as RTs, pupillary responses and blinks have been shown to measure cognitive processing (Seymour, Baker, & Gaunt, 2013) and may be useful also in measuring face recognition. These will be discussed in turn below. Our secondary aim was to investigate whether any pupillary changes were associated with fluctuations in cognitive load, or fluctuations in cognitive engagement, and to evaluate both accounts in terms of face processing.

First, we tested accuracy, but for these analyses, we excluded own face data, as all participants correctly responded to their own face, which was only shown once. We therefore only analysed data from the familiar and unfamiliar faces. To our surprise, we found that participants were more accurate in responding to unfamiliar faces than familiar faces, since previous research suggests that people are more accurate when processing familiar faces (Ellis et al. 1979; Bruce et al., 2001).

However, this could be an artefact of decision bias, which is in favour of judging faces as 'unfamiliar' at the expense of 'familiar' judgements. This bias was investigated in two experiments, the first by (Jenkins et al., 2011), who showed participants 40 faces of two Dutch celebrities. They found that people who did not know the celebrities tended to categorise the faces as being many different individuals, while people who did know the celebrities realised that the images contained different views of two individuals. The second experiment was by Bindemann and Sandford (2011). They tested people's ability to match three ID cards (each containing a different photo of one person) to the correct person among 30 photographs of different people. They found that people often thought that the ID cards displayed photos of three different people. Therefore, both experiments suggest that people have a tendency to regard different views of the same unfamiliar face as belonging to different individuals.

The analyses in the current study also revealed that accuracy was moderated by condition, and this interaction was moderated by university. The results showed that performance was poorest in the *unaware* condition, significantly so compared with the *consented* condition: it was only in the *unaware* group that participants from both universities performed less accurately when responding to familiar faces. This suggests that the unexpected appearance of their own image, negatively affected their performance when responding to the familiar faces. Kent participants who were *aware* were also more accurate when processing unfamiliar faces, but Sussex participants who were *aware* showed no such difference in accuracy. Finally, Sussex participants were more accurate in processing familiar faces in the *consented* condition, but there was no difference in accuracy in this condition in the Kent participants. These findings partially explain the less accurate familiar face results overall, and suggest that there might either have been subtle differences in the ways in which the researchers from Kent and Sussex conducted the experiment, or that the images from Sussex were easier to process.

The remaining analyses included data from the own faces, as well as the familiar and unfamiliar faces. The results of the reaction time analyses revealed only that familiar faces were processed more quickly than unfamiliar faces, supporting some previous research (Keyes & Zalicks, 2016), although there are conflicting findings (Ramon, Caharel, & Rossion, 2011). In short, reaction times were not useful for investigating own-face processing and provided no insight into the other variables. Therefore, we analysed the physiological responses to face processing, starting with pupillary responses.

Pupil sizes were largest when participants viewed their own face, although there were no significant differences in pupil size between familiar and unfamiliar faces. This suggested that the own faces elicited some kind of cognitive response in participants that

was not found in the familiar and unfamiliar faces, indicating that own faces are treated differently. When looking at the number of fixations, we found that the number of fixations decreased as faces became more familiar, supporting previous fixation research (Barton, Radcliffe, Cherkasova, Edelman, & Intriligator, 2006; Heisz & Shore, 2008), and suggesting that own faces may be understood as the most familiar of the face types. The combined results of the pupillary and fixation data suggest that these physiological responses may be useful in understanding differences in processing different face types, especially own faces. It also seems that they provide more insight into the processes than behavioural responses, as they reflect implicit processes, and are thus less likely to be affected by conscious decisions.

Having established that pupillary and fixation responses were associated with different face types, we asked whether either of the two theoretical constructs (cognitive load and cognitive engagement) could account for the responses. Starting with pupillary responses, fluctuations in cognitive load did not initially appear to account for the pupillary changes while viewing the three face types. Assuming that unfamiliar faces processing is more difficult than familiar faces processing, cognitive load would be expected to be largest while processing the unfamiliar faces, resulting in larger pupil sizes, but in this experiment the own faces produced the largest pupil sizes.

However, as the own faces were presented veridically, an account in terms of cognitive load is possible. Faces are not entirely symmetrical, so when they are mirror-reversed the asymmetry is also reversed. Until recently, when people saw their own face, it was usually in the mirror, so that they would be accustomed to seeing their facial asymmetry reversed. When seeing a photograph of their face, it would no longer be mirror-reversed, causing the asymmetry to be more apparent (Seyama & Nagayama,

2006), briefly giving the face the appearance of an unfamiliar face. As the own faces were not mirror-reversed in this experiment, not only might the familiar configurations have been disrupted by seeing the image, briefly making them appear unfamiliar, but participants also had to mentally left-right reverse the image to process it as their own face, requiring greater cognitive load than when processing the other faces. An alternative explanation could stem from the different ways in which own faces and familiar faces are processed: previous research has suggested that own faces are processed in a piecemeal fashion, potentially requiring greater mental effort than processing other familiar faces, which are processed holistically (Brédart, 2003).

Nevertheless, these accounts fail to explain the lack of interaction between face type and condition, as participants from all conditions had larger pupil sizes for their own face than for the other faces, whether they were explicitly told that they would see their face or not. This would indicate that knowing that one would see one's own face did not reduce the mental effort required to process it.

The recent popularity in taking selfies with mobile phones also makes these accounts seem less plausible. Nowadays, many people see frequent images of their own face on mobile phones or on social media, so they have an intimate knowledge of their own (unreversed) face from multiple angles as well as seeing their face in a mirror (Brédart, 2003). In short, they are more familiar with their own face than they are with any other face, whether or not its image is mirror-reversed, meaning that a cognitive load account of the large own face pupil size seems unlikely. Indeed, Wen and Kawabata (2014) found that participant's bias for 'attractive' versions for their own face among morphed and original images was not affected by mirror-reversal, and that participants did not even notice that the image was reversed, suggesting that mirror reversal does not

affect own-face recognition. It appears that the research investigating own-face processing remains inconclusive regarding the extent to which own faces appear familiar, and what effect this might have on cognitive load.

An account in terms of cognitive engagement seems more plausible. Pupil sizes were largest when participants viewed their own face (and there was no interaction between face type and condition), suggesting that their own face was more engaging than the other faces, regardless of whether or not the participant expected to see it (Proulx et al., 2017). This supports previous research that suggests that self-relevant stimuli are important, including own faces (Kircher et al., 2000; Tacikowski & Nowicka, 2010). This interest in one's own image is demonstrated in the selfie phenomenon, where doctored images are shared as idealised representations of the self (Murray, 2015), and in the tendency for people to think more 'attractive' morphs of their faces are unmodified images, an effect that disappears when presented with images of friends (Wen & Kawabata, 2014). It seems likely that engagement with images of one's own face could override fluctuations in cognitive load associated with processing faces of different degrees of familiarity.

However, there is an alternative explanation for the pupillary changes, as pupils are larger when memory strength is greater (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012). They also appear to reflect the experience of recognition (Otero et al., 2011), and may also reflect the strength of evidence on which recognition is based (Montefinese, Vinson, & Ambrosini, 2018). Therefore, the larger pupil sizes when looking at own faces may have reflected the greater memory strength associated with own faces compared to other faces.

Moving onto the fixation data, initially, the findings also appeared to support an account in terms of cognitive load: it could be argued that the number of fixations decreased as faces became more familiar, because the amount of mental effort required diminished. One explanation may be that people can rely on holistic processing when processing familiar faces, which can sometimes be achieved with just one fixation, but require more piecemeal analysis to process unfamiliar faces, requiring a greater number of fixations. However, research suggests that both unfamiliar and familiar face processing use a combination of configural and piecemeal processing (Collishaw & Hole, 2000), so this seems unlikely. The account would also contradict the own-face pupillary data, as if the pupil sizes were large due to either the image briefly appearing unfamiliar, or the cognitive demands of piecemeal processing, then participants' own faces would also be expected to produce more fixations rather than fewer. Finally, cognitive load theory fails to account for the lack of interaction between *face type* and *condition*. One would expect participants who knew that they would see their own face to have fewer fixations than those who did not, as they had been given a clue that would reduce the mental effort required for processing, but this was not the case.

Cognitive engagement may also offer a more plausible account for the fixation data, as own faces were associated with the least number of fixations (Barton et al., 2006), suggesting that participants were staring at their own face more than at unfamiliar faces. It also offers a more plausible account for the lack of interaction between conditions, which suggests that regardless of the information provided to participants about being shown their own face, participants tended to stare at it more than at other faces. This is supported in the RT analysis, as although there were no differences in the amount of time that participants looked at the different face types, there were fewer fixations when they

looked at their own face, suggesting that more of the time looking at it was spent staring at it compared with the other face types.

In short, while the present study failed to find differences in pupil size between familiar and unfamiliar faces, it appears that fixation and pupillary data are useful in understanding cognitive processes involved in face processing. Pupillary data indicated that own faces are processed differently from personally-familiar and unfamiliar faces, and fixation data indicated that the number of fixations required to process faces decreases as faces become more familiar. Both measures also provided some insight into the accounts of face processing provided by cognitive load and cognitive engagement. While no definitive conclusion could be drawn, the evidence leans towards a cognitive engagement account of the physiological changes that occurred, although memory strength also offers a plausible explanation for the pupillary responses. The pupillary and fixation data appeared to be more useful than behavioural data, as they measured implicit processes that did not appear to be dependent on conscious decisions, indicating that they could that they could prove beneficial in applied contexts such as forensic settings.

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043–1047.

Ayres, P., & Paas, F. (2012). Cognitive Load Theory: New directions and challenges. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of

familiarity, inversion, and morphing on scanning fixations. *Perception*, *35*(8), 1089–1105. https://doi.org/10.1068/p5547

Berggren, N., Koster, E. H. W., & Derakshan, N. (2012). The effect of cognitive load in emotional attention and trait anxiety: An eye movement study. *Journal of Cognitive Psychology*, *24*(1), 79–91. https://doi.org/10.1080/20445911.2011.618450

Binda, P., Pereverzeva, M., & Murray, S. O. (2014). Pupil size reflects the focus of feature-based attention. *Journal of Neurophysiology*, *112*(12), 3046–3052. https://doi.org/10.1152/jn.00502.2014

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625–627. https://doi.org/10.1068/p7008

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brédart, S. (2003). Recognising the usual orientation of one's own face: The role of asymmetrically located details. *Perception*, *32*(7), 805–811. https://doi.org/10.1068/p3354

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037//1076-898X.7.3.207

Buetti, S., & Lleras, A. (2016). Distractibility is a function of engagement, not task difficulty: Evidence from a new oculomotor capture paradigm. *Journal of Experimental Psychology: General*, *145*(10), 1382–1405. https://doi.org/10.1037/xge0000213

Burton, A.M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces: Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Carbon, C.C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801–806. https://doi.org/10.1068/p5789

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Collishaw, S. M., & Hole, G. J. (2000). Featural and configurational processes in the recognition of faces of different familiarity. *Perception*, *29*(8), 893–909. https://doi.org/10.1068/p2949

Devue, C., Van der Stigchel, S., Brédart, S., & Theeuwes, J. (2009). You do not find your own face faster; you just look at it longer. *Cognition*, *111*(1), 114–122. https://doi.org/10.1016/j.cognition.2009.01.003

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431–439.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Heisz, J., & Shore, D. (2008). More efficient scanning for familiar faces. *Journal of Vision, 8*(1), 9.1-10.

Jainta, S., & Baccino, T. (2010). Analysing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Jenkins, R., White, D., Van Montfort, X., & Burton, M.A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnston, P. R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596. https://doi.org/10.1080/09658210902976969

Keyes, H., & Zalicks, C. (2016). Socially important faces are processed preferentially to other familiar and unfamiliar faces in a priming task across a range of viewpoints. *PLOS ONE*, *11*(5), e0156350. https://doi.org/10.1371/journal.pone.0156350

Kircher, T. T. J., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P. J., Bullmore, E. T., … David, A. S. (2001). Recognizing one's own face. *Cognition*, *78*(1), B1–B15. https://doi.org/10.1016/S0010-0277(00)00104-9

Kosaka, H., Omori, M., Iidaka, T., Murata, T., Shimoyama, T., Okada, T., … Wada, Y. (2003). Neural substrates participating in acquisition of facial familiarity: an fMRI study. *NeuroImage*, *20*(3), 1734–1742. https://doi.org/10.1016/S1053-8119(03)00447-6

Laeng, B., & Falkenberg, L. (2007). Women's pupillary responses to sexually significant others during the hormonal cycle. *Hormones and Behavior*, *52*(4), 520–530. https://doi.org/10.1016/j.yhbeh.2007.07.013

Laurence, S., & Hole, G. (2011). The effect of familiarity on face adaptation. *Perception*, *40*(4), 450–463. https://doi.org/10.1068/p6774

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77–100. https://doi.org/10.1037/0096-1523.34.1.77

Lorini, E., & Castelfranchi, C. (2007). The cognitive structure of surprise: looking for basic principles. *Topoi*, *26*(1), 133–149.

Mathôt, S., Siebold, A., Donk, M., & Vitu, F. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, *144*(3), 513–521. https://doi.org/10.1037/a0039168

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037//1076-8971.7.1.3

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251–274.

Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: the pupil's point of view. *Biological psychology*, *135*, 159-169.

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, *23*(5), 1316-1340. https://doi.org/10.3758/s13423-015-0982-5

Murray, D. C. (2015). Notes to self: the visual culture of selfies in the age of social media. *Consumption Markets & Culture*, *18*(6), 490–516. https://doi.org/10.1080/10253866.2015.1052967

Ninomiya, H., Onitsuka, T., Chen, C.-H., Sato, E., & Tashiro, N. (1998). P300 in response to the subject's own face. *Psychiatry and Clinical Neurosciences*, *52*(5), 519–522. https://doi.org/10.1046/j.1440-1819.1998.00445.x

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory: Pupil size and recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Proulx, T., Sleegers, W., & Tritt, S. M. (2017). The expectancy bias: Expectancy-violating faces evoke earlier pupillary dilation than neutral or negative faces. *Journal of Experimental Social Psychology*, *70*, 69-79.

Pilz, K. S., Bülthoff, H. H., & Vuong, Q. C. (2009). Learning influences the encoding of static and dynamic faces and their recognition across different spatial frequencies. *Visual Cognition*, *17*(5), 716–735. https://doi.org/10.1080/13506280802340588

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Prehn, K., Heekeren, H. R., & Van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an

analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Ramasubbu, R., Masalovich, S., Gaxiola, I., Peltier, S., Holtzheimer, P. E., Heim, C., … Mayberg, H. S. (2011). Differential neural activity and connectivity for processing one's own face: A preliminary report. *Psychiatry Research: Neuroimaging*, *194*(2), 130–140. https://doi.org/10.1016/j.pscychresns.2011.07.002

Ramon, M., Caharel, S., & Rossion, B. (2011). The speed of recognition of personally familiar faces. *Perception*, *40*(4), 437–449. https://doi.org/10.1068/p6794

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: a PET study of face categorical perception. *Journal of Cognitive Neuroscience*, *13*(7), 1019–1034.

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage*, *37*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Seyama, J., & Nagayama, R. S. (2006). Can mirroring reveal image distortion? Illusory distortion induced by mirroring. *Psychological Research Psychologische Forschung*, *70*(2), 143–150. https://doi.org/10.1007/s00426-004-0189-2

Seymour T. L., Baker C. A., Gaunt J. T. (2013). Combining blink, pupil and response time measures in a concealed knowledge test. *Frontiers Psychol, 3*: 1–15.

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

Tacikowski, P., & Nowicka, A. (2010). Allocation of attention to self-name and self-face: An ERP study. *Biological Psychology*, *84*(2), 318–324. https://doi.org/10.1016/j.biopsycho.2010.03.009

Tong, F., & Nakayama, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1016–1035. https://doi.org/http://dx.doi.org/10.1037/0096-1523.25.4.1016

Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E., & Iacoboni, M. (2005). Self-face recognition activates a frontoparietal 'mirror' network in the right hemisphere: an event-related fMRI study. *NeuroImage*, *25*(3), 926–935. https://doi.org/10.1016/j.neuroimage.2004.12.018

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wen, W., & Kawabata, H. (2014). Why am I not photogenic? Differences in face memory for the self and others. *I-Perception*, *5*(3), 176–187. https://doi.org/10.1068/i0634

Wu, E. X. W., Laeng, B., & Magnussen, S. (2012). Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition. *Social Neuroscience*, *7*(2), 202–216. https://doi.org/10.1080/17470919.2011.596946

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

CHAPTER 4. YOU CAN BELIEVE YOUR EYES: MEASURING IMPLICIT RECOGNITION IN A LINEUP WITH PUPILLOMETRY.

Abstract

As pupil size is affected by cognitive processes, we investigated whether it could serve as an independent indicator of target recognition in lineups. Participants saw a simulated crime video, followed by two viewings of either a target-present or target-absent video lineup while pupil size was measured with an eye-tracker. For participants who made correct identifications, pupil sizes were significantly larger when viewing the target compared with distractors. Also, some participants were uncertain about their choice of face from the lineup, but nevertheless showed pupillary changes when viewing the target, suggesting that there had been covert recognition of the target face. The results suggest that pupillometry might be a useful aid in assessing the accuracy of an eyewitness' identification.

Keywords: pupillometry, eyewitness identification, covert recognition, face processing

It has long been understood that recognising familiar faces is accomplished with a high degree of accuracy, whilst recognising unfamiliar faces is more problematic (see Hancock, Bruce, & Burton, 2000, for a review). For example, in a study using CCTV images and comparison photographs, Bruce, Henderson, Newman and Burton (2001) reported that matching accuracy was about 75% for unfamiliar faces, compared to approximately 90% for familiar faces. Eyewitnesses are required to recognise individuals often seen only briefly before, despite this task being extremely difficult (Hancock et al. 2000). There is considerable evidence that the inaccurate identifications made by eyewitnesses are a major factor in miscarriages of justice (the Innocence Project, n.d.; see also Dwyer, Neufeld, & Scheck, 2000; Wells & Olson, 2003).

Given the unreliability of eyewitnesses' responses to lineups, and the fact that police are unable to differentiate between recognition and non-recognition on the basis of identification responses alone, research has investigated ways in which to assess the credibility of eyewitnesses via other means (e.g. MacLin, MacLin, & Malpass, 2001; Wright & Stroud, 2002). Using ERPs (event-related potentials), Lefebvre, Marchand, Smith and Connolly (2007) found that participants who made correct identifications showed an increased P300 response to the target compared to distractors. The P300 was also significantly larger for correct identifications than for misidentifications. However, although this is a promising result, it would currently be impractical to measure ERPs in a real-world setting.

In an attempt to compare scores on established face recognition tests with lineup responses, Bindemann, Brown, Koyas and Russ (2012) found that a face recognition test postdicted lineup performance in participants who made an identification, but not in participants who misidentified the target, missed the target, or correctly rejected all faces

in a target-absent condition. In a follow-up study, the face recognition test only postdicted lineup performance for correct rejections. Both experiments (across target-present and target-absent lineups) indicated that the face recognition test provided a good index of eyewitness reliability for participants who made an identification, but not for those who made no identification.

Confidence has also been studied with regards to whether it can be diagnostic of identification accuracy. There is evidence that judges and juries attach considerable weight to a witness' confidence when evaluating them (e.g. Wells, Ferguson, & Lindsay, 1981). Although not all research has found confidence ratings to be a reliable guide to identification accuracy (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988), more recent research suggests that the confidence-accuracy relationship is stronger than previously thought (Wixted, Read, & Lindsay, 2016). If confidence can be recorded immediately after an identification is made, including before any feedback is provided, then there is evidence that it can be a useful indicator in instances where the witness makes a selection from the lineup (Sauerland & Sporer, 2009; Sauer, Brewer, Zweck, & Weber, 2009). In addition, research has shown that very confident eyewitnesses tend to have relatively high degrees of accuracy whereas the same is not true of unconfident witnesses (Brewer & Palmer, 2010). Both the American Psychology-Law Society (Wells et al., 1998) and the US National Research Council (2014) recommend, therefore, that confidence should be recorded immediately following a lineup. However, one problem with confidence ratings is that they are susceptible to being influenced by post-identification feedback, especially if made retrospectively rather than immediately after the lineup has taken place (Wells & Bradfield, 1998).

Another measure that has been used in eyewitness research (e.g. Sauerland & Sporer, 2009; Meissner, Tredoux, Parker, & MacLin, 2005) is the remember-know (RK) paradigm. It was introduced to measure states of awareness associated with memory retrieval, and originally used to differentiate likely accuracy in semantic memory (Conway & Dewhurst, 1995). It has subsequently been refined by many researchers to also include "Guess" responses, for instance, when participants are not certain when looking at an 'old' stimulus but do not want to select 'new' (see Dunn, 2004, for a review).

The use of ERPs, confidence and remember/know judgements is aimed at providing information about cognition, particularly the accuracy of memory, without requiring the conscious input of the participant. Another potential method of achieving the same goal is to use pupillometry. Pupillometry is potentially useful because research has shown that pupil size is not determined solely by ambient luminance, but can be influenced by cognitive load: the greater the mental workload, the larger the pupil size (Beatty, 1982; Jainta & Baccino, 2010; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; & see Ayres & Paas, 2012; Goldinger & Papesh, 2012; Murphy, Groeger, & Greene, 2016, for reviews). Pupil size has also been associated with affective processing: pupils are larger when presented with emotional stimuli than with neutral stimuli (e.g. Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016).

Pupillometry has also proved useful in indexing memory strength, as pupils have been shown to be larger when retrieving items associated with greater memory strength (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012), and they also appear to reflect the experience of

recognition (Otero et al., 2011). Therefore, they may also reflect the strength of recognition evidence (Montefinese, Vinson, & Ambrosini, 2018).

Pupillometry also appears to be useful for measuring implicit memory, as pupillary changes occur in the absence of an overt response (van Rijn, Dalenberg, Borst, & Sprenger, 2012), and can even occur despite efforts to deceive. For instance, Heaver and Hutton, (2011) found that pupil sizes were larger when looking at words that had been previously seen in a list, compared to new words. This was despite giving different instructions to participants, either to feign memory loss or to perform as accurately as possible. Thus, pupil size reflected memory strength that was independent of the overt responses that participants gave.

Pupillometry has seldom been used in face recognition research. However, Goldinger, He and Papesh (2009) have shown that pupil sizes were larger when looking at other-race faces than own-race faces. Considering the social importance of faces, and combining this with the findings that pupils respond to memory strength, it appears that pupil changes could be a reliable measure of face recognition, and applicable to eyewitness lineup procedures (review in Goldinger & Papesh, 2012). However, as far as we know, pupil size has not been investigated within the context of a lineup until now.

Pupillometry has the potential to be a useful supplementary measure of eyewitness identification performance. It would be desirable to have a measure of eyewitness performance that is independent of the explicit decision processes involved in making an identification. The previous attempts to assess eyewitness accuracy, using measures of witness confidence or generalised face recognition ability, do not fulfil this criterion: these are alternative explicit measures of recognition that may well be contaminated by the conscious decision processes involved in making an identification

in the first place. In contrast, pupil size is a physiological response outside of the witness' conscious control, and hence more likely to be independent of their overt decision about the lineup. Unlike traditional measures, pupillometry also reveals fluctuations in cognitive processing, such as changes in mental effort, engagement and memory strength *at the time of viewing a suspect's face*. This suggests that pupillary measures may provide more nuanced information about eyewitness identification performance, which could be used to assist decisions about an eyewitness' credibility.

The present study investigated whether pupil size is a good predictor of lineup identification accuracy. Pupillary responses were compared to two overt measures of identification performance: the witness' identification from the lineup, and their assessment of the strength of their memory for the face they had seen (using "Remember", "Know" and "Guess" decisions, e.g. Johnson & Wellman, 1980). It was predicted that pupillary changes would reflect memory strength. Specifically, it was predicted that pupillary changes would be largest in participants who successfully identified the target because they remembered their face: in these participants, pupil sizes would be larger when viewing the target compared with viewing distractors. The study was conducted with both target-present and target-absent lineups. Target-present lineups were employed so as to be able to determine whether pupil size changes can be used to identify when a participant is viewing a face they have seen previously (i.e. to discriminate between the target face and distractors). However, target-absent lineups were also employed, as it is important to know what would happen to pupil size if the perpetrator of the crime was *not* present in the lineup. To be a potentially effective method, pupillometry would need to consistently differentiate target faces from distractors in target-present lineups *and* show that pupil size changes are not associated consistently with any single face in the lineup, especially when the target is absent.

4.2. Methods

4.2.1. Participants

In the target-present condition, 51 participants with normal or corrected-to-normal vision were recruited at the University of Sussex in exchange for course credits or cash. Two participants were subsequently excluded because they made multiple identifications in both lineups, contrary to the experimenter's instructions. This left 49 participants for the analysis (15 males and 34 females). They were aged between 18 and 26 (M = 19.61, SD = 1.72). Participants were recruited until there were at least ten for each category of identification response (identifiers, non-identifiers and misidentifiers) in each lineup presentation. In the target-absent condition there were 26 participants (2 males, 24 females), aged between 18 and 35 (M = 20.42, SD = 3.23). This study was approved by the Sciences & Technology Cross-Schools Research Ethics Committee (crecscitec@sussex.ac.uk). The project reference number is ER/CE214/5.

4.2.2. Apparatus and Stimuli

In the familiarisation stage, participants saw a silent video clip of a staged non-violent crime, in which a man attempted to steal another man's bag. After a brief altercation, the two men ran off, still fighting over the bag. The video lasted one minute. It was recorded in .wmv format (768 x 576 pixels) and converted to XVID for compatibility with the eye-tracking software, Experiment Builder (SR Research, n.d.).

Two types of lineup were used: target-present and target-absent. In the target-present condition, the lineup stage involved sequential presentation of 10 colour video clips of head and shoulders against a white background (nine distractors and one target). In each clip, the individual initially faced the camera. Then they turned their head slowly

to the right, back to centre, to the left and back to centre. The video clips were constructed by the VIPER Unit of the West Yorkshire Police, using the VIPER (n.d) video identification parade database. Trained officers selected distractors from the VIPER database (containing thousands of faces) to match a basic verbal description of the target (whilst ensuring that all distractors were also a reasonable visual match to the physical appearance of the target). Thus, the faces were matched as closely as possible in terms of age, race, attractiveness and so on. Videos were cropped and matched for size (17.5 x 13.3 cm), resolution (768 x 576 pixels), time (12 seconds), and luminance. The blinds were drawn to control the room's lighting levels. In the target-absent condition, the target face was removed from the lineup, so participants only saw the 9 distractor faces.

Experiment Builder was run using a 21.5 inch iMac computer and a desktop Eyelink 1000 eye-tracker (SR Research, n.d.) that recorded pupil position and size using infrared illumination. The participant's head was stabilised with a chin rest, at an approximate distance of 60 cm from the computer screen that displayed the stimuli. The right eye was tracked for all participants.

4.2.3. Design

This study used a mixed design: independent measures on identification response (with three levels: identifiers, non-identifiers, and misidentifiers) and repeated measures on face type (with three levels: pre-target faces, target face, and post-target faces). With target-present lineups, the target face was the person seen in the staged crime video. With target-absent lineups, the "target" face was the misidentified lineup member who the witnesses had not encountered before the lineup took place. This procedure is described in detail in section 3.2. The dependent variable was pupil size, calculated as a percentage of each participant’s overall pupil size range during the experiment (see 4.3. for details).

4.2.4. Procedure

Participants were briefed before placing their chins in the chin rest, and their eye movements were calibrated to nine points on the computer screen. After reading instructions on the screen, their gaze was monitored with a drift check. Drift checks monitor the accuracy of eye-tracking data and involve looking at a black dot on a white screen. At this point the video clip of the simulated crime was played. This was followed by a filler task, which was the long version of the Glasgow Face Matching Task (GFMT) (Burton et al., 2010). The task was chosen to reflect the fact that eyewitnesses see many faces between a real crime and the lineup, so we considered it to be more ecologically valid than isolating participants from any faces between the crime video and the lineup in the experiment. The GFMT task took between 8 and 33 minutes per participant (M = 17.79, SD = 4.74).

Immediately after the filler task, and before seeing the lineup, participants were given practice at identifying a face: they were shown two faces that did not resemble the target. These were filmed in the same way as the other lineup faces, and participants were told explicitly that they were for practice only. (A pilot version of this experiment revealed that participants who misidentified a distractor tended to do so when looking at the first face in the lineup, suggesting that they would benefit from becoming acquainted with the task before viewing the lineup itself).

After completing the practice session, participants saw a hybrid video lineup (similar to those used in UK police procedures) relating to the staged crime. Lineup video clips were displayed one at a time with a drift check between adjacent clips. For each clip, the participant was asked to click ‘Y’ if they thought the face was the target and ‘N’ if they thought it was a distractor. They were asked to respond as quickly and accurately

as possible for each clip and asked not to press ‘Y’ more than once per lineup. Each clip played for 12 seconds regardless of how quickly the participant responded, but the ISI was not fixed, as each clip was separated by a drift check. This took approximately 2-3 seconds: as soon as their eye stabilised, the next trial began. Each participant saw the lineup video clips in a different pseudo-random order (in the case of the target-present lineups, the target was never first or last in the lineup). After the final clip, the procedure was repeated, with the clips displayed in a different pseudo-random sequence, and participants were asked to make Yes/No responses as they had for the first lineup presentation. They were told that this response could be the same as it had been in the first lineup presentation, or that they could make a different response if they wished. The eye-tracker recorded eye movement data and responses as participants viewed the clips. Following the task, participants were asked to rate their memory strength during their identification performance using a version of the RKG paradigm (Appendix 1.).

4.3. Results

To standardise pupil size measurements between participants, the following procedure was used. For each participant, the eye-tracker produced a mean pupil size for each video clip in the lineup. We subtracted each participant's smallest mean from their largest mean, to produce a difference score. The mean pupil size for each clip was then expressed as a percentage of this difference.

From these values, in the target-present condition, three pupil size measures were then produced for each participant. First, pupil sizes for distractors seen *before* the target were averaged together to produce a single mean pupil size measure, labelled “pre-target distractors”. There was only one target, so only one measure was available for each

participant, labelled “target”. Finally, pupil sizes for distractors seen *after* the target were averaged together to give a single mean pupil size measure labelled “post-target distractors”. A similar method was used in the target-absent condition, where the erroneously identified distractor (for participants that made an identification) replaced the actual target. This procedure is described in detail in section 4.3.2.

4.3.1. Target-Present condition.

In order to determine whether pupil sizes changed in response to the target face, two two-way mixed ANOVAs were conducted, one for the first lineup and another for the second. For each ANOVA, there was one within-subjects factor, *face type* (with three levels: pre-target distractors, target, and post-target distractors) and one between-subjects factor, *identification response* (with three levels: identifiers, non-identifiers, and misidentifiers).

Participants were divided into three categories based on their identification response: "identifiers", participants who correctly identified the target; "non-identifiers", participants who mistakenly thought the target was absent; and "misidentifiers", participants who mistook a distractor for the target. Some participants made multiple misidentifications in one of the lineups but not the other. Their data were only analysed for the lineup in which they performed as instructed. This left 45 participants who were included in the analysis for the first lineup presentation.

Bayes factors are useful for assessing the strength of evidence of a theory, and for drawing different conclusions from those of orthodox statistical methods. Orthodox statistics model the null hypothesis (H_0), revealing whether there is a statistical difference between means, but nothing else. Bayes factors make three-way distinctions: whether the

data either supports the null hypothesis (H_0); whether they strengthen support for the alternative hypothesis (H_1); or whether there is no evidence either way. They also challenge perceptions of the importance of power, as they indicate that a high-powered non-significant result is not always evidence to support the H_0, but a low-powered non-significant result might be. Similarly, a high-powered significant result might not be substantial evidence of H_1. Finally, using Bayes one can specify the hypothesis in a way that is not possible with a *p* value (Dienes & McLatchie, 2017). Therefore, Bayes Factors were also calculated for key non-significant results in the current experiment (Singh, n.d.). A pilot experiment indicated that the mean difference in pupil sizes for the target compared with distractors was 14.39% (in the first lineup). Therefore, the *SD* was set to $x = 14.39$ when making the same comparisons in the present experiments. The mean difference in pupil sizes between identifiers and participants who did not identify the target was 9.60%, so the *SD* for this comparison was set to $x = 9.60$.

Inspection of fig. 49 suggests that, for the first lineup presentation, pupil sizes were larger when viewing the target compared to distractors, and different between the three identification response groups (identifiers, non-identifiers and misidentifiers). This interpretation was supported by the ANOVA results for the first lineup, which showed significant main effects of *face type*, F (2,84) = 25.55, $p < .001$, $r = .48$, $\eta^2 = .38$ and *identification response*, F (2,42) = 3.75, $p = .03$, $r = .29$, $\eta^2 = .15$, but no significant interaction between *face type* and *identification response*, F (4,84) = 2.27, $p = .07$.

Using the target as a baseline, planned contrasts revealed that pupil sizes in participants who correctly identified the target in the first lineup were significantly larger (28.4%) when viewing the target than when viewing pre-target distractors, F (1,21) = 82.30, $p < .001$, $r = .89$ and post-target distractors F (1,21) = 99.22, $p < .001$, $r = .91$

(32%). In participants who made no identification, pupils were significantly larger (21%) when viewing the target than when viewing post-target distractors, F (1,12) = 9.99, p = .01, r = .67, but not when viewing the pre-target distractors F (1,12) = 1.79, p = .21 (9.4%), although the Bayes factor B_H = 1.60 indicated that the results were insensitive. There were no significant comparisons in participants who misidentified a distractor: pre-target distractors, F (1,9) = 1.05, p = .10 (8%), although the Bayes factor B_H = 1.17 indicated that the results were insensitive; post-target distractors F (1,9) = 4.33, p = .07 (20%). However, the Bayes factor B_H = 4.73 supported the alternative hypothesis.

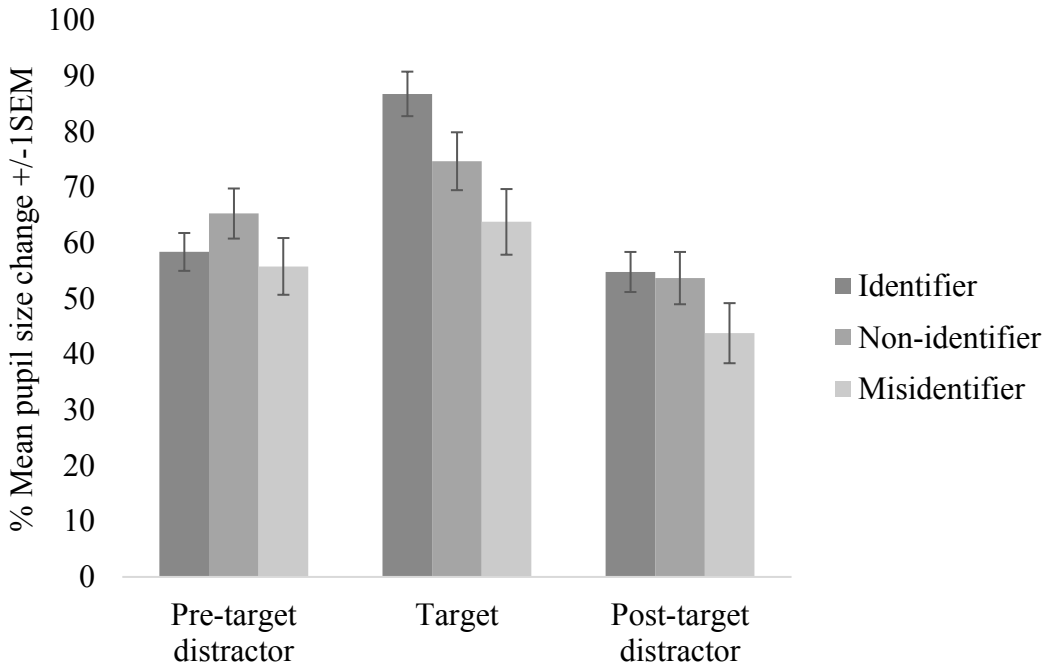


Fig. 49. Pupillary changes in response to the first lineup presentation:

(Legend:) Pupillary changes for pre-target distractors, target, and post-target distractors in the first lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

We conducted three one-way ANOVAs to compare the three groups of participants for each type of face. Identifiers, non-identifiers and misidentifiers showed significant pupillary differences when looking at the target, F (2,44) = 5.51, p = .01, r = .33, but not for pre-target distractors, F (2,44) = 1.16, p = .33, and the Bayes factor B_H = 0.14 also indicated that the results supported the null hypothesis; or post-target distractors, F (2,44) = 1.52, p = .23, although the Bayes factor B_H = 1.35 indicated that the results were insensitive.

For the second lineup presentation, data from 44 participants were analysed. There was a significant main effect of *face type*, F (2,82) = 12.28, p < .001, r = .36, η^2 = .23 (Pre: M = 41.70, SE = 2.70; Target: M = 51.40, SE = 3.90; Post: M = 32.70, SE = 2.90), no effect of *identification response*, F (2,41) = 0.25, p = .77, (Identifiers: M = 42.60, SE = 3.20; Misidentifiers: M = 39.40, SE = 4.70; Non-identifiers: M = 43.80, SE = 4.30 and the Bayes factor B_H = 0.12 supported the null hypothesis). There was no interaction between *face type* and *identification response*, F (4,82) = 0.92, p = .46. As can be seen from fig. 50, overall pupil sizes were larger when viewing the target than distractors. Thus, pupillary responses discriminated between identifiers, misidentifiers and non-identifiers the first time they saw the lineup, but not when they viewed it a second time.

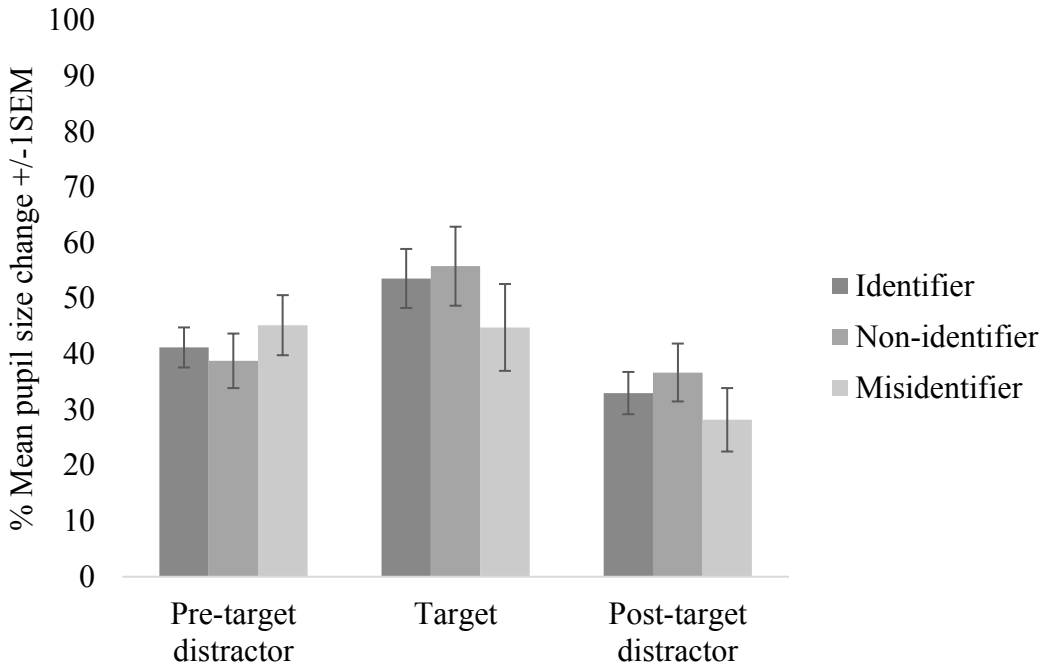


Fig. 50. Pupillary changes in response to the second lineup presentation:

(Legend:) Pupillary changes for pre-target distractors, target, and post-target distractors in the second lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

4.3.1.1. Using pupil size to predict identification response:

Two binary logistic regressions were used to determine whether pupil size change could predict whether participants made a correct or incorrect lineup decision. The predictor variable was *pupil size* (calculated as the mean difference between the target and the distractors) and the outcome variable was *identification accuracy* (correct or incorrect).

For the first lineup, the logistic regression was statistically significant, $\chi^2(1) = 6.49$ $p = .01$. The model explained 17.6% (Nagelkerke R^2) of the variance in lineup

decision outcome (i.e. whether or not participants were correct in their decision) and correctly classified 69.6% of cases. For the first lineup presentation, *pupil size* was therefore a fairly good measure of *identification performance*. This was not true for the second lineup, for which the logistic regression was not significant, $\chi^2(1) = 0.50$, $p = .48$.

4.3.1.2. Participants' subjective assessments of identification accuracy:

Two Chi Square analyses were used to see whether participants' assessment of their 'memory strength' (in terms of their "Remember", "Know" or "Guess" responses) was related to their actual performance with the lineup. There was no significant association between memory strength and performance in either lineup presentation: first lineup presentation, $\chi^2(2) = 1.27$, $p = .53$; second lineup presentation, $\chi^2(2) = 1.13$, $p = .57$.

Next, we investigated whether pupillary changes were related to the RKG responses, taking into account whether or not the witness made a correct identification. Two three-way ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with two levels: target, and distractors) and two between-subjects factors, *identification accuracy* (with two levels: correct and incorrect), and *RKG rating* ("Remember", "Know" or "Guess"). The dependent variable was pupil size change.

For the first lineup presentation, there was a significant main effect of *face type*, $F(1,42) = 43.71$, $p < .001$, $r = .70$, $\eta^2 = .51$. There were also interactions between *face type* and *accuracy*, $F(1,42) = 8.89$, $p = .01$, $\eta^2 = .18$ and between *face type*, *accuracy* and *RKG*, $F(2,42) = 3.68$, $p = .03$, $\eta^2 = .15$. However, there was no interaction between *face type* and *RKG*, $F(2,42) = 1.89$, $p = .16$.

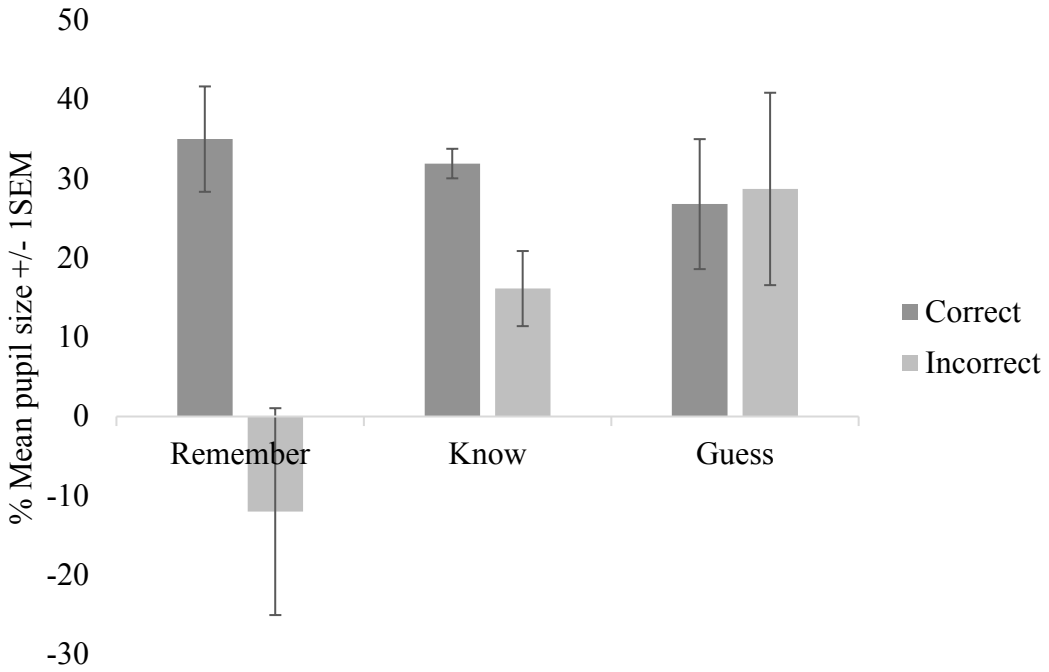


Fig. 51. Mean pupillary difference between the target and distractors in the first lineup presentation, as a function of identification accuracy (correct and incorrect), and RKG rating (remember, know and guess).

(Legend:) Positive: mean pupil size was larger when looking at the target than at the distractors. Negative: mean pupil size was smaller when looking at the target than at the distractors.

As seen in fig. 51, pupil size did not differ for correct and incorrect participants who guessed. Both groups had pupillary responses to the target although the error bars were large, indicating that pupillary responses were not good measures of identification response in these participants. For the "Know" responders, pupil size changes in response to the target were consistent with explicit identification decisions when participants were correct, but not when they were incorrect. However, for the "Remember" responders, pupil size changes were strikingly consistent with the identification response: pupil sizes were 35% larger when looking at the target compared to the distractors. However, in

those who failed to identify the target despite saying that they remembered him, pupil sizes were 12% smaller than when looking at the distractors.

In participants who rated their memory strength as “Know”, pupil sizes were 32% larger when looking at the target compared to the distractors. However, in those who failed to correctly identify the target despite claiming that they recognised him, pupil sizes were 16% larger in response to the target face than when looking at the distractors.

For the second lineup presentation, there was a significant main effect of *face ty*pe, F (1,41) = 13.87, $p < .001$, $r = .49$ (Pre: $M = 41.70$, $SE = 2.70$, Target: $M = 51.40$, $SE = 3.90$, Post: $M = 32.70$, $SE = 2.90$), but no other significant effects (largest $F = 0.93$).

4.3.2. Target-Absent condition

Participants were divided into two categories based on their identification response: "misidentifiers", participants who mistook a distractor for the target, and “correct rejectors”, those who correctly responded that the target was not present in the lineup.

Three pupil size measures were taken from each participant. In participants who misidentified a distractor, in the absence of a target we wanted to see pupillary responses to the face that was misidentified, so we treated this face as the "target", but called it the "false positive". Therefore, pupil sizes for distractors seen before the false positive were averaged together to produce a single mean pupil size measure, labelled “pre-false positive distractors (Pre)”. There was only one false positive, so only one measure was available for each participant, labelled “false positive”. Finally, pupil sizes for distractors seen after the false positive were averaged together to give a single mean pupil size measure labelled “post-false positive distractors (Post)”. In participants who correctly

rejected all faces, we did not even have a false positive face to compare to the target. Therefore, we selected the distractor that had been misidentified most often (47% of the time) and designate this face to be the "false positive", as we considered it most likely to be considered a close match to the target and therefore most likely to elicit a large pupil size. Then, we followed the same procedure that we had followed for the target-absent misidentifiers.

To test whether pupil sizes changed in response to a misidentified face in the absence of the target, two two-way mixed ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with three levels: pre, false positive, and post) and one between-subjects factor: *identification response* (with two levels: correct rejectors and misidentifiers).

As seen in fig. 52, for the first lineup presentation there was a significant main effect of *face type*, F (2,48) = 5.92, p = .01, r = .44, η^2 = .20 (Pre: M = 61.00, SE = 3.80, False positive: M = 69.40, SE = 4.60, Post: M = 50.20, SE = 3.90), but there was no effect of *identification response*, F (1,24) = 0.80, p = .38 (Misidentifiers: M = 62.30, SE = 4.00; Correct rejectors: M = 57.70, SE = 3.20), although the Bayes factor B_H = 1.01 indicated that the results were insensitive. There was also no interaction between *face type* and *identification response*, F (2, 48) = 0.36, p = .70. While misidentified faces (and faces likely to be misidentified) elicit larger pupil sizes than other faces, pupillary responses did not discriminate between misidentifiers and correct rejectors.

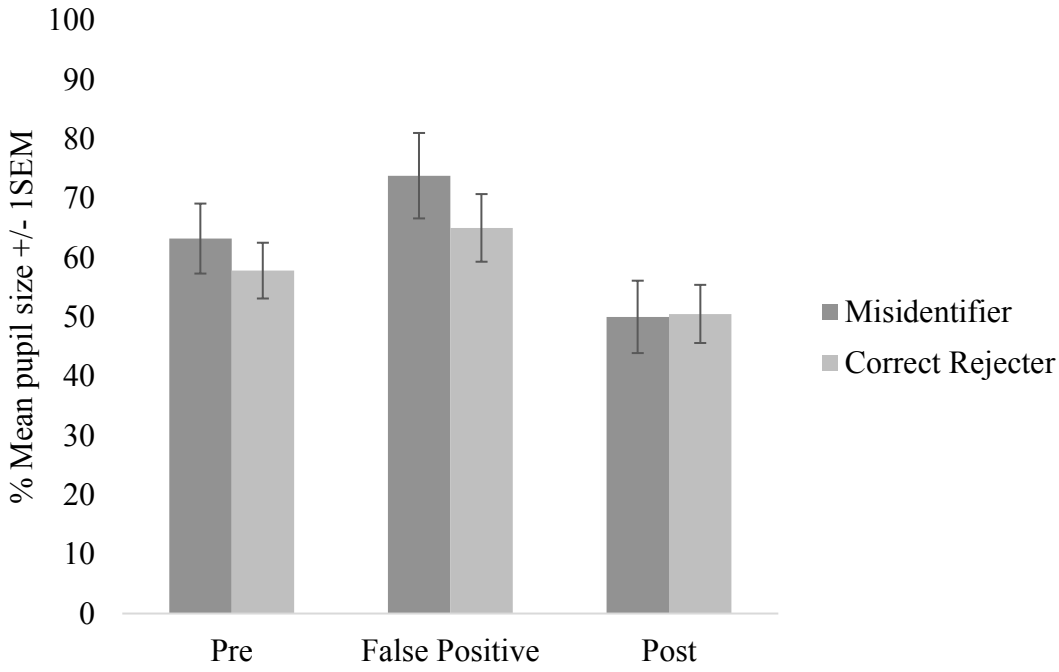


Fig. 52. Pupillary changes in response to the first lineup presentation:

(Legend:) Pupillary changes for pre-false positive distractors, false positive, and post-false positive distractors in the first lineup presentation, with participants grouped by identification response (correct rejectors and misidentifiers).

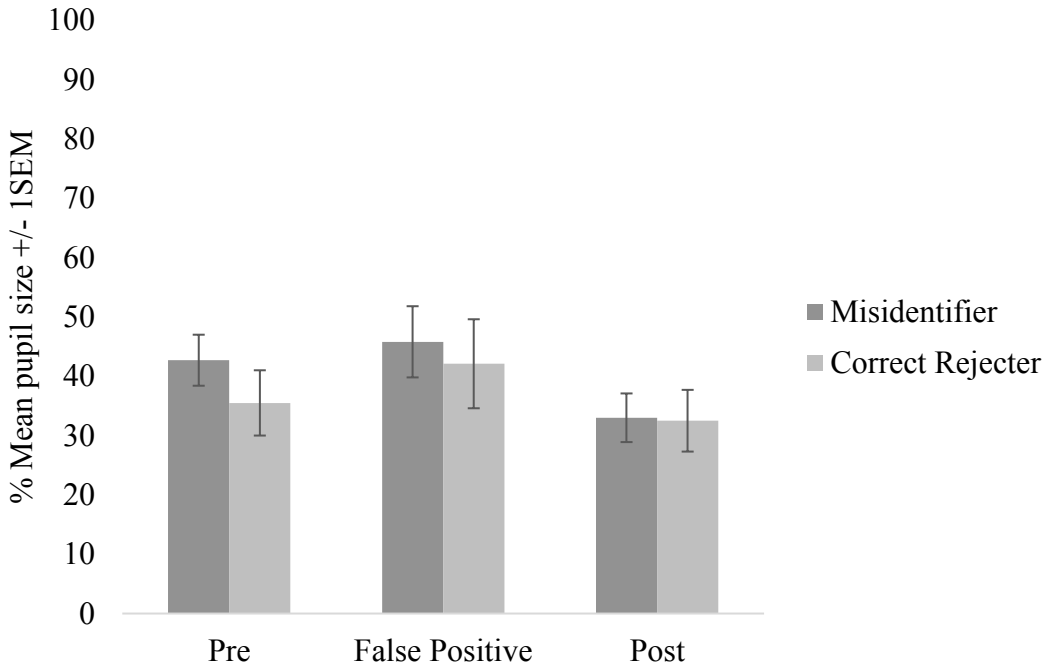


Fig. 53. Pupillary changes in response to the second lineup presentation:

(Legend:) Pupillary changes for pre-false positive distractors, false positive, and post-false positive distractors in the second lineup presentation, with participants grouped by identification response (correct rejectors and misidentifiers).

As seen in fig. 53, for the second lineup there were no significant effects *face type* F (2,48) = 2.61, p = .08 (Pre: M = 39.10, SE = 3.50; False positive: M = 43.90, SE = 4.80; Post: M = 32.80, SE = 3.30), although the Bayes factor B_H = 8.31 supported the alternative hypothesis. There was also no significant effect of *identification response*, F (1,24) = 0.50, p = .49 (Misidentifiers: M = 40.50, SE = 3.40; Correct rejectors: M = 36.70, SE = 4.20), although the Bayes factor B_H = 0.86 indicated that the results were insensitive. Finally, there was no interaction between *face type* and *identification response*, F (2, 48) = 0.23, p = .80. While pupil sizes were not *significantly* larger when viewing misidentified faces (and faces likely to be misidentified), Bayesian analysis supported the alternative

hypothesis, and indicated that more participants were required to determine whether pupillary changes were different between correct rejectors and misidentifiers.

4.3.2.1. Using pupil size to predict identification response:

Two binary logistic regressions were used to determine whether pupil size change could predict whether participants made a correct or incorrect lineup decision. The predictor variable was *pupil size* (calculated as the mean difference between the target and the distractors) and the outcome variable was *identification accuracy* (correct or incorrect).

The logistic regression was not statistically significant for either lineup presentation: first lineup presentation, $\chi^2(1) = 1.20$, $p = .27$; second lineup presentation, $\chi^2(1) = 0.23$, $p = .88$. Pupil size did not predict whether participants would correctly reject all the faces or misidentify a face.

4.3.2.2. Participants' subjective assessments of identification accuracy:

Two Chi Square analyses were used to see whether participants' own assessment of their 'memory strength' was related to their actual lineup performance. There was no significant association between memory strength and performance for either lineup presentation: first lineup presentation, $\chi^2\ (2) = 3.00$, $p = .22$; second lineup presentation, $\chi^2\ (2) = 4.76$, $p = .09$. Therefore, participants' assessment of their memory was not a good indicator of their performance.

Next, we investigated whether pupillary changes were related to the RKG responses, taking into account whether or not the witness made a correct identification. Two three-way ANOVAs were used for analysis of pupil size measures in response to

each lineup. For each, there was one within-subjects factor, *face type* (with two levels: false-positive, and distractors) and two between-subjects factors: *identification accuracy* (with two levels: correct and incorrect), and *RKG rating* ("Remember", "Know" and "Guess"). The dependent variable was pupil size change.

For the first lineup presentation, there was a significant main effect of *face type*, F (1,21) = 10.51, p = .01, r = .58, but no other significant effects (largest F = 3.19). For the second lineup presentation, there was a marginal effect of *face type*, F (1,21) = 3.96, p = .06, but no other effects (largest F = 1.39).

4.4. Discussion

The principal finding of this study is that pupil size changed in response to the target in target-present lineups. This only occurred in participants who correctly identified the target, and only the first time that they saw him. These pupillary responses also predicted identification of the target the first time participants saw him. In regard to memory strength, RKG responses had no bearing on identification accuracy, but pupillary changes were related to the RKG responses when participants were divided according to their identification accuracy. In contrast, in the target-absent condition pupillary responses did not differentiate between those who correctly rejected the faces and those who misidentified a face. This suggested that it was only the presence of a previously-seen face that resulted in the pupillary effects in the target-present condition. RKG responses were not related to identification accuracy or pupillary responses.

This novel approach to measuring implicit recognition in a lineup with pupillometry is in line with previous research suggesting that pupillary changes are associated with memory strength (Otero et al., 2011; Papesh et al., 2012; Brocher & Graf,

2016). Our results indicate that when people recognised the target in a lineup, their pupils became larger in response to his face. However, this was only found for the first lineup presentation. In the second lineup presentation all the faces had been seen before, so it may be that different cognitive requirements, such as source monitoring (being able to place the correct face to the crime) had also affected pupil sizes. The Bayes Factors also indicated that more participants were needed in some of the analyses, so testing more people might clarify the results somewhat.

In light of our findings, it was important to assess whether pupil size was also able to predict identification of the target, and we found that it did. In the first lineup presentation, pupil size change explained just under 18% of the variance in lineup decision outcome and correctly classified just under 70% of cases. As expected from the initial analyses, in the second lineup pupil size was not a good predictor of identification accuracy.

Another measure we wanted to investigate was the RKG paradigm. Confidence scales have been used more widely in eyewitness research, and are used in US legal proceedings (Sauerland & Sporer, 2009; Sauer et al., 2009; US National Research Council; 2014). However, as pupillometry appears to be a good measure of memory strength, we considered the RKG paradigm to be more appropriate here (Dunn, 2004). First, we wanted to see whether RKG responses reflected participants' identification performance. We found that participants appeared to have little insight into their performance, as RKG responses had no bearing on their identification performance.

However, as RKG is a measure of a person's belief in their memory strength, and memory strength is measured by pupil size, we also wanted to see whether pupillary changes were related to the explicit RKG responses. We tested this by taking into account

whether or not the participant had made a correct identification. These analyses were very revealing. For the first lineup presentation it was clear that pupillary changes were related to RKG responses. Specifically, they showed that patterns of pupillary changes differed considerably when participants with different assessments of their memory strength were correct or incorrect. Fig. 58 showed that when participants said they "guessed", pupillary responses did not help to differentiate correct and incorrect participants any more than their identification responses did. However, when participants "remembered" or "knew" the target's face, pupillary responses were better at distinguishing between correct and incorrect participants. The results from these participants indicated that when participants rated their memory as strong, their pupils also indicated that their memory strength was strong, but only when the participants were *correct*. When people rated their memory as strong and were *incorrect*, there was no discernible trace of memory as measured by the pupil size. Again, there were no significant effects in the second lineup presentation, suggesting that it is only when the target is the only familiar face in the lineup that the effect occurs.

Having established that pupillary responses to the target occurred in participants who identified him, and that pupillary responses were related to RKG responses when we divided participants on the basis of their identification of the target, we ran the same analyses for a target-absent condition. As expected, pupils did not respond as they had done in the target-present condition: they did not reflect explicit identification responses, they did not predict the explicit identification responses, and they were not related to the explicit RKG responses. Thus, we concluded that the pupillary changes that had occurred in the target-present condition had done so specifically in response to the target, and reflected memory strength for the target's face.

Interestingly, the pupillary changes for the second lineup presentation in the target-present condition appeared to be as indistinct as those of the target-absent condition. This suggests that the pupillary changes that occur when viewing the target only occur when the distractors are novel (in the first lineup presentation). Once all the faces have been seen, the familiarity distinction between the target and the distractors disappears, meaning that the pupillary changes also disappear. It seems to be the distinction between entirely-novel faces and a previously-seen face that elicits pupillary changes in participants who remember that face.

We wanted to assess pupillometry in relation to traditional measures of lineup identification that rely on explicit decisions. Identification decisions require eyewitnesses to choose between identifying a face or not. This measure is supposed to be based upon their recognition of the suspect as the perpetrator. To do this, the eyewitness first needs to weigh up the options against their memory, and then make an explicit choice. However, these decisions rely on conscious processes that might mask implicit recognition in some cases, for instance if implicit recognition has not reached the threshold required for an identification, if an eyewitness does not want to make an error (for fear of wrongful conviction), or if they actually know the perpetrator and make a decision not to identify them.

Confidence ratings and RKG responses require people to judge their performance. These can be contaminated, for example by post-identification feedback (Sauerland & Sporer, 2009; Sauer et al., 2009), and may reflect traits such as self-concept (Kröner & Biermann, 2007) more than the task at hand. Like decision responses, they can also reflect the eyewitness's choices, in regard to wanting to make an identification or not. Pupillometry appears to overcome these shortcomings, as pupillary changes occur in the

absence of an overt response (van Rijn et al., 2012), and can occur despite efforts to deceive (Heaver & Hutton, 2011). Thus, pupil size appears to reflect memory strength that is independent of explicit responses.

The pupillometry data may have some relevance for police lineup procedures. In the US, law enforcement agencies tend to use simultaneous lineups (containing six faces), but some agencies use a hybrid system in which witnesses can choose whether or not to view a sequential lineup presentation for a second time. In the UK, police use a hybrid sequential system (containing nine faces), with two presentations of a sequential display (Seale-Carlisle & Mikes, 2016). The effectiveness of hybrid systems like these has been tested by Steblay, Dietrich, Ryan, Raczynski and James (2011). They found that participants picked a face more times in the second lineup presentation than the first, but more of these choices were misidentifications than identifications. Also, participants who elected to view two presentations were less accurate than those who chose to see just one, and more likely to perform worse with the second lineup presentation than they had with the first. Our research also showed that there were no benefits to having a second lineup display, as the second lineup did not produce the pupillary changes that had provided insights into memory strength in the first.

More encouragingly, our research suggested that the UK system of presenting faces sequentially would be a good choice if pupillometry was used to measure memory strength. While this has not been tested, it is anticipated that pupillometry not be able to detect fluctuations in memory strength in simultaneous presentations, as participants are able to move their gaze freely around the display. This would not give pupils time either to adjust for each face or reset between faces. It is also anticipated that some faces would

be dismissed after a mere glance or using peripheral vision, so there would be no pupillary data for them at all.

Finally, the results presented here show that pupillometry provided a measure of recognition strength that appeared to be independent of participants' identification responses, suggesting it could be a potentially important measure for improving the accuracy of eyewitness identification evidence. When the target was present, and participants make a correct identification, pupils changed in a way that reflected their explicit RKG rating. However, when participants believed that they had a strong memory for the target's face but failed to identify him, their pupils did not change. Also, when the target was absent, and the participants mistakenly responded that he was present, the lack of pupil size changes showed this was a mistake. Therefore, if an eyewitness makes an identification, but their pupils show little change or get smaller, then they are likely to be wrong. If an eyewitness makes no identification, but their pupils get larger when seeing the suspect, it suggests that they might have recognised them implicitly. Thus, pupil sizes do not just mirror the participants' overt decision processes, but provide insight into implicit memory processes.

We do not propose that pupillary responses can replace explicit responses in police lineups, but our research suggests that they can offer a measure of implicit memory strength that provides insight into the identification responses that people make. As a consequence, pupillometry could be a practical tool to support current measures, and could shed light on the strengths and weaknesses of the processes currently used.

References

Ayres, P., & Paas, F. (2012). Cognitive Load Theory: New directions and challenges. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*(3), 353-364.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology*, *15*(1), 77–96. https://doi.org/10.1348/135532509X414765

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology*, *53*(12), 1823–1835. https://doi.org/10.1111/psyp.12770

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037//1076-898X.7.3.207

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291.

Conway, M. A., & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *The Quarterly Journal of Experimental Psychology Section A*, *48*(1), 125–140. https://doi.org/10.1080/14640749508401380

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-56.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207-218.

Dunn, J. C. (2004). Remember-Know: A matter of confidence. *Psychological Review*, *111*(2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524

Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: five days to execution and other dispatches from the wrongly convicted* (1st ed). New York: Doubleday.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

The Innocence Project (n.d.), retrieved 18th May, 2018, from https://www.innocenceproject.org/causes/eyewitness-misidentification/

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Johnson, C. N., & Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know, and guess. *Child development*, 1095-1102.

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept—Towards a model of response confidence. *Intelligence*, *35*(6), 580-590.

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs).

Psychophysiology, *44*(6), 894–904. https://doi.org/10.1111/j.1469-8986.2007.00566.x

MacLin, O. H., MacLin, M. K., & Malpass, R. S. (2001). Race, arousal, attention, exposure and delay: An examination of factors moderating face recognition. *Psychology, Public Policy, and Law*, *7*(1), 134–152. https://doi.org/10.1037//1076-8971.7.1.134

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783–792. https://doi.org/10.3758/BF03193074

Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: the pupil's point of view. *Biological psychology*, *135*, 159-169.

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, *23*(5), 1316-1340. https://doi.org/10.3758/s13423-015-0982-5

National Research Council. (2015). *Identifying the culprit: Assessing eyewitness identification*. National Academies Press.

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory: Pupil size and recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Prehn, K., Heekeren, H. R., & van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2009). The effect of retention interval on the Confidence–Accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*, *3*(9), 160300. https://doi.org/10.1098/rsos.160300

Singh, A.K. (n.d.), Bayes Factor (Dienes) calculator, retrieved 5th September, 2018, from https://medstats.github.io/bayesfactor.html

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration,

habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

SR Research (n.d.), retrieved, April 23rd, 2018, from https://www.sr-research.com/products/eyelink-1000-plus/

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262–274. https://doi.org/10.1007/s10979-010-9236-2

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS ONE*, *7*(12), e51134. https://doi.org/10.1371/journal.pone.0051134

VIPER (n.d.), retrieved January 16th, 2018, from http://www.viper.police.uk

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360-376.

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*(6), 688.

Wells, G. L., Small, M., Penrod, S. J., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, 22, 603-647.

Wells, G. L., & Olson, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wixted, J. T., Read, D. J., & Lindsay, S. D. (2016). The effect of retention interval on the eyewitness identification Confidence–Accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *5*(2), 192–203. https://doi.org/10.1016/j.jarmac.2016.04.006

Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: people are better with their own age. *Law and Human Behavior*, *26*(6), 641.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

CHAPTER 5. POLICING POSITIVE IDENTIFICATIONS: MEASURING IMPLICIT RECOGNITION IN POLICE LINEUPS WITH PUPILLOMETRY.

Abstract

Eyewitness identification responses can result in wrongful convictions and non-convictions. Research has investigated why eyewitnesses make identification mistakes, how procedures can influence identification responses, and measured eyewitness credibility. Traditional behavioural responses can be contaminated by conscious decisions, and fail to index 'live' recognition of specific faces. Pupillometry has been shown to measure implicit recognition in lineups and is independent of identification responses. However, no research has investigated this in a UK hybrid video lineup procedure. We recorded pupillary responses with an eye-tracker as participants viewed either a target-present or target-absent lineup. We found that pupil sizes changed in participants who identified the target, when they looked at his face. The results were evaluated theoretically, and they provided a means for assessing current police systems. They also suggest that pupillometry could be a practical tool for assisting with credibility assessments in UK police procedures.

Inaccurate eyewitness identification responses can result in miscarriages of justice: the wrongful conviction of an innocent person, that can occur when a distractor is misidentified, or when the suspect is identified by chance although the police have the wrong suspect; and the non-conviction of a guilty person, that can occur when a distractor is misidentified or when nobody is identified in the lineup. The first can have devastating consequences for an innocent person who is convicted, and the second can put people at risk from dangerous offenders. Therefore, the burden placed on eyewitnesses is considerable, yet eyewitness identifications often fail to be accurate because the task is so difficult (the Innocence Project, n.d.; see also Dwyer, Neufeld, & Scheck, 2000; Wells & Olson, 2003). This is because the perpetrator's face is unfamiliar to them, and recognising unfamiliar faces is not easy (see Hancock, Bruce, & Burton, 2000, for a review).

Therefore, eyewitness research has investigated why eyewitness identification is so error-prone, by looking into variables associated with the eyewitness or the event that can affect identification, such as race or age. This is because it has been found that people are more likely to misidentify a face of another race or age (e.g. Wells & Olson, 2001; Wright & Stroud, 2002; Memon, Bartlett, Rose, & Gray, 2003; Havard & Memon, 2009; Havard, Memon, Laybourn, & Cunningham, 2012; Wylie, Bergt, Haby, Brank, & Bornstein, 2015). Other variables include expectations (Allport & Postman, 1947) intoxication (Yuille & Tollestrup, 1990; Yuille, Tollestrup, Marxsen, Porter, & Herve, 1998; Hagsand, Hjelmsäter, Granhag, Fahlke, & Söderpalm-Gordh, 2013), stress (e.g. Valentine & Mesout, 2009; Rush et al., 2014), and exposure time or delays between witnessing a face and viewing it in a lineup (Loftus, Schooler, Boone, & Klein, 1987; Read, 1995; & see MacLin, MacLin, & Malpass, 2001 for a review), all factors which have been shown to affect the likelihood of a correct identification.

However, error-prone witnesses are only part of the problem. Another key issue is determining whether the responses that eyewitnesses give are correct or not. The police have to determine whether an eyewitness who identifies the suspect did so because they recognised them as the perpetrator, or whether the suspect was selected merely by chance (even though they were not the perpetrator). They also have to determine whether an eyewitness who makes *no identification* did so because the suspect was not the perpetrator, or because the eyewitness failed to recognise them as such. However, all the police have to help them make these determinations are identification responses, which are fairly unreliable measures of recognition. There is considerable evidence that inaccurate identifications are a major factor in miscarriages of justice (the Innocence Project, n.d.; see also Dwyer et al., 2000; Wells & Olson, 2003).

Given this evidence, and the fact that police cannot differentiate between recognition and non-recognition on the basis of identification responses alone, researchers have investigated the credibility of eyewitnesses via other measures (e.g. MacLin et al., 2001; Wright & Stroud, 2002) such as eyewitness confidence. Judges and juries have been found to take into account an eyewitness' confidence when evaluating credibility (e.g. Wells, Ferguson, & Lindsay, 1981), but not all research has found that confidence is a satisfactory measure of identification accuracy (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988). As a result, confidence ratings are not used in UK Police lineups. However, more recent research suggests that the confidence-accuracy relationship is more reliable than previously thought (Wixted, Read, & Lindsay, 2016). For instance, it can be a useful indicator of accuracy if recorded immediately after an identification is made, including before any feedback is provided (Sauerland & Sporer, 2009; Sauer, Brewer, Zweck, & Weber, 2009). Therefore, both the National Academy of

Sciences (2014) and the American Psychology-Law Society (Wells et al., 1998) recommend that confidence should be recorded immediately after a lineup.

Another measure that has been used in eyewitness research (e.g. Sauerland & Sporer, 2009; Meissner, Tredoux, Parker, & MacLin, 2005) is the remember-know (RK) paradigm, which also asks participants to rate their performance. It relates to eyewitness recognition as it was introduced to measure states of awareness associated with memory retrieval (Conway & Dewhurst, 1995). Many researchers have since included include "Guess" responses for instances where “Remember” and “Know” are insufficient (see Dunn, 2004, for a review).

Bindemann, Brown, Koyas and Russ (2012) attempted to test eyewitness credibility by comparing scores on established face recognition tests with lineup responses. This test has the potential to predict lineup performance on the basis on general recognition ability, rather than relying on self-ratings of lineup performance. They found that a face recognition test was a reliable measure of eyewitness performance when participants made an identification, but not for those who made no identification.

Lefebvre, Marchand, Smith and Connolly (2007) used event-related potentials (ERPs) to measure physiological responses to recognition that did not rely on self-ratings or general recognition ability, and found that participants who correctly identified the target showed an increased P300 response to the target compared to distractors. The advantage of ERPs over the other methods is that they provide physiological measures that are probably independent of decision responses, but it would currently be impractical to measure ERPs in forensic settings.

Another potential method of obtaining physiological responses more practically is to use pupillometry. Pupillometry is potentially useful because pupil size can be influenced by cognitive processes like cognitive load, as with greater mental workloads pupil sizes get larger (Beatty, 1982; Jainta & Baccino, 2010; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; & see Ayres & Paas, 2012; Goldinger & Papesh, 2012; Murphy, Groeger, & Greene, 2016, for reviews). Pupil size is also influenced by the emotional content of the stimuli: pupils get larger when the stimuli are emotional (e.g. Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016).

Pupils have been shown to be larger when retrieving items associated with greater memory strength (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012), and can appear to reflect the experience of recognition (Otero et al., 2011) or strength of evidence (Montefinese, Vinson, & Ambrosini, 2018). Pupillary changes even occur without an overt response (van Rijn, Dalenberg, Borst, & Sprenger, 2012), and can even occur despite efforts to deceive. Despite giving different instructions to participants, either to feign memory loss or to perform accurately, Heaver and Hutton (2011) found that pupil sizes were larger when looking at 'old' words compared to 'new' words. Thus, pupil size reflected memory strength that was independent of participants' overt responses.

Pupillometry has seldom been used in face recognition research (review in Goldinger & Papesh, 2012), although Goldinger, He and Papesh (2009) have shown that pupil sizes were larger when looking at other-race faces than own-race faces. More recently, we conducted a study that investigated whether pupillometry could provide a measure of implicit recognition, *while faces were presented* during a hybrid lineup with

two sequential video lineup presentations (Chapter 4). We found that pupil sizes changed in response to the target in target-present lineups. This only occurred in participants who correctly identified him, and only the first time they saw him. Pupil size thus provided a fairly good predictor of identification response in the first lineup presentation. The results also showed that pupillometry provided a measure of implicit recognition strength that was independent of conscious identification responses.

Our research demonstrated the potential use of pupillometry as an additional tool for assessing eyewitness performance in the first presentation of a sequential display, suggesting that it could be useful in police displays that are presented sequentially. There are currently variations of two approved methods of presenting faces to eyewitnesses: simultaneous and sequential. Simultaneous lineups show photographic images of all the faces at the same time, while sequential lineups show images (either photographs or videos) one at a time. Research suggests that people make more correct identifications in simultaneous lineups than in sequential lineups, but at the expense of also making more misidentifications. This is because the simultaneous method encourages them to make relative judgments, so they tend to pick the face that is the best fit to their memory of the perpetrator (Flowe & Cottrell, 2011). Sequential lineups reduce misidentifications because they encourage absolute judgments, which means that the face being assessed is compared directly to the memory of the perpetrator (Cutler and Penrod 1988; Lindsay and Wells 1985; Sporer 1993). However, neither procedure is entirely satisfactory as both only produce about 25% correct identifications overall (Wells, Steblay, & Dysart, 2015).

Therefore, in the UK, police use a hybrid sequential system (containing nine faces), where there are two presentations of a sequential display (Seale-Carlisle & Mickes, 2016). Most law enforcement agencies in the US tend to use the simultaneous

system containing six faces (Seale-Carlisle & Mickes, 2016), but a hybrid system was also introduced in the US, where the eyewitnesses have an *option* to see the lineup presentation twice. It was tested by Steblay, Dietrich, Ryan, Raczynski & James (2011), who found that the second lineup elicited more identifications than the first. They also found that participants who *chose* to have two presentations were less accurate than those who did not, and were also more likely to perform worse in the second lineup presentation. Recent research with over 2000 participants suggests that the American system of simultaneous displays is most reliable (Seale-Carlisle & Mickes, 2016). However, the many differences between the UK and US systems, and conflicting results from multiple studies indicate that more research is warranted.

In the present research we wanted to see whether pupillometry would benefit current UK police procedures. We chose the UK procedure on the basis that pupillometry has already been shown to reflect memory strength in a hybrid video lineup (Chapter 4) and UK lineup parades use video lineups. However, their research used a procedure that required the participants to respond to each face (with a Yes/No response). Therefore, in the present research, we used a procedure where the faces were numbered, in line with current police methods. This requires participants to remember the number of the face that they wish to identify.

However, we also used four additional measures. We asked participants to make an identification response once for each lineup, so that we could compare accuracy between lineup presentations. We asked participants to rate their memory strength with a remember-know-guess (RKG) paradigm after each lineup presentation, so that we could see whether participants had any insight into their memory strength, and whether this changed between lineup presentations. We included a practice session, as our research

suggested that this improved both accuracy with the actual lineup and the reliability of the pupillary data (Chapter 4). Finally, we asked participants to type the number of the face they were looking at, so that we could be sure that they had seen the number.

The main reason for doing this experiment was to see whether pupillometry could provide support in current UK procedures, that require eyewitnesses to memorise the number of the face that they wish to identify. We did not know whether having to remember numbers allocated to the faces would incur an additional memory load to providing Yes/No responses to each face, and thus affect pupillary responses.

On the assumption that it would support our previous research (Chapter 4), we predicted that pupillometry would provide a measure of implicit recognition. In the target-present condition, participants who recognised the target in the first lineup presentation would have larger pupils when looking at his face compared to those of the distractors, and pupillary responses would predict whether a participant identified the target or not. Pupillary responses in participants who "Remembered" the target in the first lineup presentation would be associated with their identification response, but this would not be the case in participants who claimed to "Guess". In the target-absent condition, we expected pupillary responses to be similar across faces and lineup presentations, both in participants who correctly rejected the faces and those who misidentified a distractor.

5.2. Methods

5.2.1. Participants

In the target-present condition, 66 participants were recruited from the University of Sussex in exchange for cash or course credits. Four were removed due to technical

issues, leaving sixty-two participants (9 males and 53 females) aged between eighteen and thirty-five (M = 20.60, SD = 3.19). Participants were recruited until there were at least ten for each category of identification response (identifiers, non-identifiers and misidentifiers) in each lineup presentation. In the target-absent condition there were 22 participants (5 males, 16 females, and one participant who did not give their gender), aged between 18 and 40 (M = 22.19, SD = 6.04). This study was approved by the Sciences & Technology Cross-Schools Research Ethics Committee (crecscitec@sussex.ac.uk). The project reference number is ER/CE214/5.

5.2.2. Apparatus and Materials

Participants first saw a silent video clip of a mock non-violent crime, in which a man tried to steal another man's bag. The video lasted one minute. It was recorded in .wmv format (768 x 576 pixels) and converted to XVID for compatibility with the eye-tracking software, Experiment Builder (SR Research, n.d.).

Two versions of a lineup were used: target-present and target-absent. In the target-present condition, the lineup stage involved sequential presentation of nine colour video clips of head and shoulders against a white background (eight distractors and one target). Each individual initially faced the camera. Then they turned their head to the right, back to centre, to the left and back to centre. Each video clip was also assigned a number from 1-9. The number was displayed clearly at the top left of the screen throughout each video clip. In the target-absent lineup, participants only saw the eight distractors. The video clips were constructed by the VIPER Unit of the West Yorkshire Police, using the VIPER (n.d.) video identification parade database. Trained officers selected distractors from thousands of faces within the VIPER database, to match both a verbal description of the target and a reasonable visual match to the physical appearance of the target. Videos were

cropped and matched for size (17.5 x 13.3 cm), resolution (768 x 576 pixels), time (12 seconds), and luminance. The room's lighting levels were controlled by drawing the blinds. Experiment Builder was run on a 21.5 inch Apple computer and an EyeLink 1000 eye-tracker, which uses an infrared camera. It stabilizes the head using a chin rest, that was set at an approximate distance of 60 cm from the computer screen that displayed the stimuli.

5.2.3. Design

This study used a mixed design: independent measures on *identification response* (with three levels: identifiers, non-identifiers, and misidentifiers) and repeated measures on *face type* (with three levels: pre-target faces, target face, and post-target faces). With target-present lineups, the target face was the person seen in the staged crime video. With target-absent lineups, the "target" face was the misidentified lineup member who the witnesses had not encountered before the lineup took place. This procedure is described in detail in section 5.3.2. The dependent variable was pupil size, calculated as a percentage of each participant's overall pupil size range during the experiment (see 5.3. for details).

5.2.4. Procedure

Participants were briefed before we calibrated their eye movements to nine points on the computer screen. Their gaze was monitored with a drift check that involved looking at a black dot on a white screen. (This helped to maintain eye-tracking accuracy during the task). After this, a video clip of a simulated crime was played. This was followed by the short version of the Glasgow Face Matching Task (GFMT) (Burton et

al., 2010). This was chosen as a filler task to reflect the fact that eyewitnesses are exposed to faces between a crime and a lineup.

After the GFMT, participants were instructed that they would see a lineup relating to the mock crime that they had just seen. This was followed by a practice session, where they had the opportunity to see how the lineup faces and numbers would be displayed. In this practice session, participants were shown two faces that did not resemble the target, but were filmed in the same way as the lineup faces. Participants' attention was drawn to the number that was allocated to each face. They were asked to type the allocated number on the computer keyboard, to ensure that they had seen it. After the practice session, participants were given further instructions on the screen that advised them that they would now see a video lineup, in which the perpetrator might or might not be present. Participants were also told that if they thought they recognised one of the faces as that of the perpetrator, they should remember its allocated number, as they would be asked for this at the end of the presentation. These instructions led them to a drift check, followed by the video lineup that included the face of the target and eight distractor faces in the target-present condition, or just the eight distractor faces in the target-absent condition. This meant that there were different numbers of faces between lineup conditions. This decision was taken to reduce confounding variables by introducing a new face in the target-absent lineup that had not been seen in the target-present lineup.

Each video of a single face in the lineup was played for 12 seconds. The ISI was not fixed, but each clip was separated by a drift check that took approximately 2-3 seconds. The lineup video clips were presented in a different pseudo-random order (the target was never one of the first two or last two faces in the lineup). After the final clip, participants provided their response: if they wanted to make an identification they typed

the number of the face that they thought was the perpetrator, and if they did not want to make an identification they typed '0'. After this, they were asked to provide the RKG rating for their performance ("Remember", "Know", "Guess"). In the second lineup presentation, the clips were displayed in a different pseudo-random sequence, and participants were asked to make a new response. They were told that their response could be the same as it had been in the first lineup presentation, or that they could change their mind if they wished. The eye-tracker recorded eye movement data and responses throughout.

5.3. Results

The eye-tracker recorded a mean pupil size for each face. To standardise pupil size measurements between participants, the following procedure was used. For each participant, we converted this to a percentage of their pupil size change during the experiment, by identifying the face that elicited the largest mean pupil size and the face that elicited the smallest mean pupil size, and calculating the difference between them. The mean pupil size for each face was then calculated as a percentage of that difference.

From these values, three pupil size measures were taken from each participant. First, pupil sizes for distractors seen *before* the target were averaged together to produce a single mean pupil size measure, labelled "pre-target distractors". There was only one target, so only one measure was available for each participant, labelled "target". Finally, pupil sizes for distractors seen *after* the target were averaged together to give a single mean pupil size measure labelled "post-target distractors".

5.3.1. Target-Present condition.

In order to determine whether pupil sizes changed in response to the target face, two two-way mixed ANOVAs were conducted, one for the first lineup and another for the second. For each ANOVA, there was one within-subjects factor, *face type* (with three levels: pre-target distractors, target, and post-target distractors) and one between-subjects factor, *identification response* (with three levels: identifiers, non-identifiers, and misidentifiers).

Participants were divided into three categories based on their identification response: "identifiers", participants who correctly identified the target; "non-identifiers", participants who mistakenly thought the target was absent; and "misidentifiers", participants who mistook a distractor for the target.

Inspection of fig. 54. suggests that, for the first lineup, there was no significant effect of *face type*, F (2,118) = 2.98, p = .06 (Pre: M = 51.90, SE = 2.70; Target: M = 59.40, SE = 3.00; Post: M = 52.20, SE = 2.30). There was a significant effect of *identification response*, F (2,59) = 4.97, p = .01, r = .28, η^2 = .20 (Identifiers: M = 61.30, SE = 2.10; Misidentifiers: M = 50.50, SE = 3.60; Non-identifiers: M = 51.70, SE = 3.40). However, there was no significant interaction between *face type* and *identification response,* F (4,118) = 1.92, p = .11.

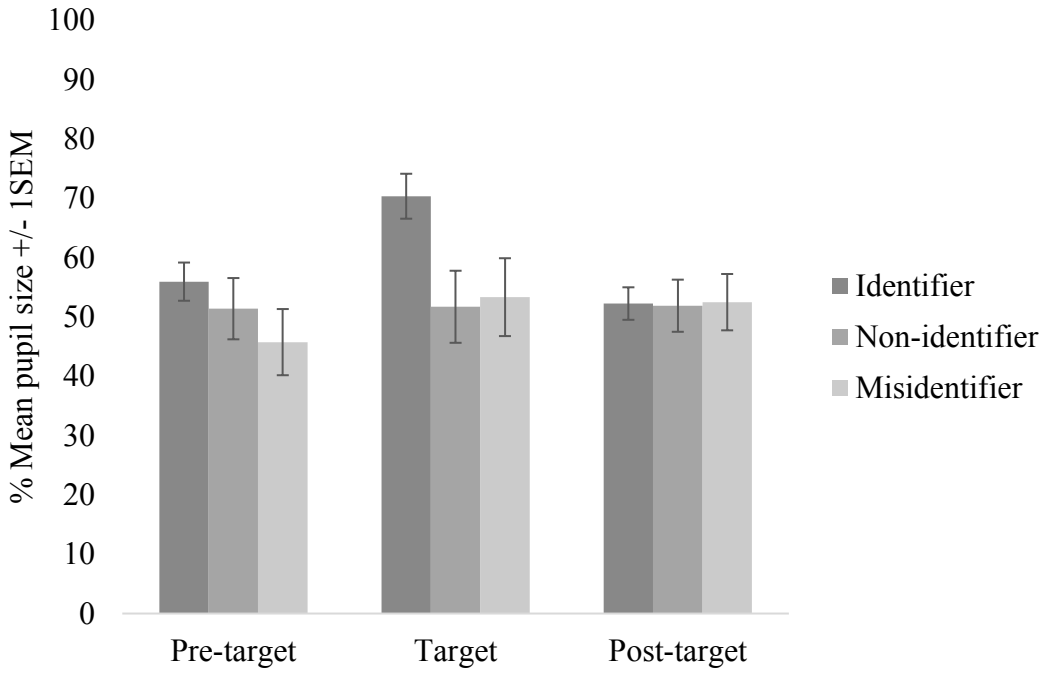


Fig. 54. Pupillary changes in response to the first lineup presentation:

(Legend:) Pupillary changes for pre-target distractors, target, and post-target distractors in the first lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

Using the target as a baseline, planned contrasts revealed that pupil sizes in participants who correctly identified the target in the first lineup were significantly larger (14.4%) when viewing the target than when viewing pre-target distractors, F (1,35) = 8.95, p = .01, r = .45, or post-target distractors, F (1,35) = 19.52, p < .001, r = .60 (18.1%). In participants who made no identification, pupils were no larger (0.3%) when viewing the target than when viewing pre-target distractors F (1,13) = 0.01, p = .98 (-0.2%) or post-target distractors, F (1,13) = 0.01, p = .97. There were also no significant comparisons in participants who misidentified a distractor: pre-target distractors, F (1,11) = 2.08, p = .18 (7.6%), post-target distractors F (1,11) = 0.01, p = .93 (0.8%).

For the second lineup presentation, there was also no significant effect of *face type*, F (2, 118) = 2.62, p = .08, and no effect of *identification response*, F (2, 59) = 1.06, p = .35, but there was a significant interaction between *face type* and *identification response*, F (4, 118) = 5.90, $p < .001$, $\eta^2 = .17$.

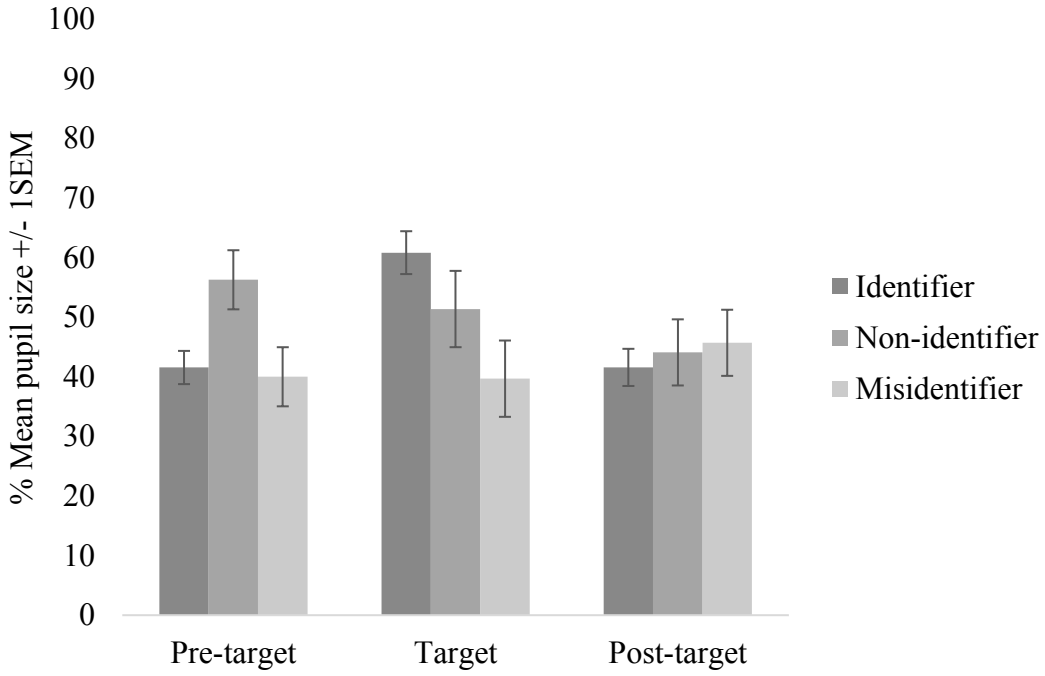


Fig. 55. Pupillary changes in response to the second lineup presentation:

(Legend:) Pupillary changes for pre-target distractors, target, and post-target distractors in the second lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

Using the target as a baseline, planned contrasts revealed that pupil sizes in participants who correctly identified the target in the second lineup were significantly larger (19.3%) when viewing the target than when viewing pre-target distractors, F (1,37) = 37.61, $p < .001$, r = .71 and post-target distractors F (1,37) = 29.36, $p < .001$, r = .67 (19.3%). In participants who made no identification, pupils were no larger (-4.9%) when

viewing the target than when viewing the pre-target distractors F (1,11) = 2.72, p = .13 (7.3%) or post-target distractors, F (1,11) = 1.11, p = .31. There were also no significant comparisons in participants who misidentified a distractor: pre-target distractors, F (1,11) = 0.01, p = .96 (-0.3%), post-target distractors F (1,11) = 0.79, p = .39 (-6%).

We conducted three one-way ANOVAs to compare the three groups of participants for each type of face. Identifiers, non-identifiers and misidentifiers showed significant pupillary differences when looking at the target, F (2,61) = 4.34, p = .02, r = .26, when looking at pre-target distractors, F (2,61) = 3.77, p = .03 r = .24, but not when looking at post-target distractors, F (2,61) = 0.24, p = .79.

Thus, pupillary responses separated correct identifiers from participants who were incorrect the *first* time they saw the lineup, and discriminated between all three groups: identifiers, misidentifiers and non-identifiers the *second* time they saw the lineup.

5.3.1.1. Using pupil size to predict identification response.

Two binary logistic regressions were used to determine whether pupil size change could predict whether participants made a correct or incorrect lineup decision. The predictor variable was *pupil size* (calculated as the mean difference between the target and the distractors) and the outcome variable was *identification accuracy* (correct or incorrect).

For the first lineup, the logistic regression model was statistically significant, $\chi^2(1)$ = 7.34, p = .01. The model explained 15% (Nagelkerke R^2) of the variance in lineup decision outcome (i.e. whether or not participants were correct in their decision) and correctly classified 67.7% of cases. For the second lineup, the logistic regression model was also statistically significant, $\chi^2(1)$ = 11.16 p = .01. The model explained 22.4%

(Nagelkerke R^2) of the variance in lineup decision outcome (i.e. whether or not participants were correct in their decision) and correctly classified 72.6% of cases. Therefore, pupil size was a fairly good measure of identification performance in both lineup presentations.

5.3.1.2. Participants' subjective assessments of their identification accuracy:

Two Chi Square analyses were used to see whether participants' assessment of their 'memory strength' (in terms of their "Remember", "Know" or "Guess" responses) was related to their actual performance with the lineup. There was a significant association between ratings of memory strength and performance in both lineup presentations: first lineup presentation, $\chi^2(2) = 6.01$, $p = .05$; second lineup presentation, $\chi^2(2) = 9.49$, $p = .01$.

Table 5. Percentage of people (and raw frequencies) in each RKG response group to identify the target correctly or not to identify him, in the first lineup presentation.

	Correct	**Wrong**
Remember	75.00% (15)	25.00% (5)
Know	56.00% (19)	44.00% (15)
Guess	25.00% (2)	75.00% (6)

Table 6. Percentage of people (and raw frequencies) in each RKG response group to identify the target correctly or not to identify him, in the second lineup presentation.

	Correct	Wrong
Remember	80.00% (16)	20.00% (4)
Know	60.00% (21)	40.00% (14)
Guess	14.00% (1)	86.00% (6)

Next, we investigated whether pupillary changes were related to the RKG responses, taking into account whether or not the witness made a correct identification. Two three-way ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with two levels: target, and distractors) and two between-subjects factors, *identification accuracy* (with two levels: correct and incorrect), and *RKG rating* ("Remember", "Know" and "Guess"). The dependent variable was pupil size change.

For the first lineup presentation, there was a significant main effect of *face type*, F (1,56) = 5.49, p = .02, r = .30. There was also an effect of *identification accuracy*, F (1,56) = 6.19, p = .02, r = .32: pupillary changes were almost 8% larger in participants who correctly identified the target than in those who did not (correct: M = 62.44, SE = 3.62; incorrect: M = 50.70, SE = 3.03), but there were no other significant effects.

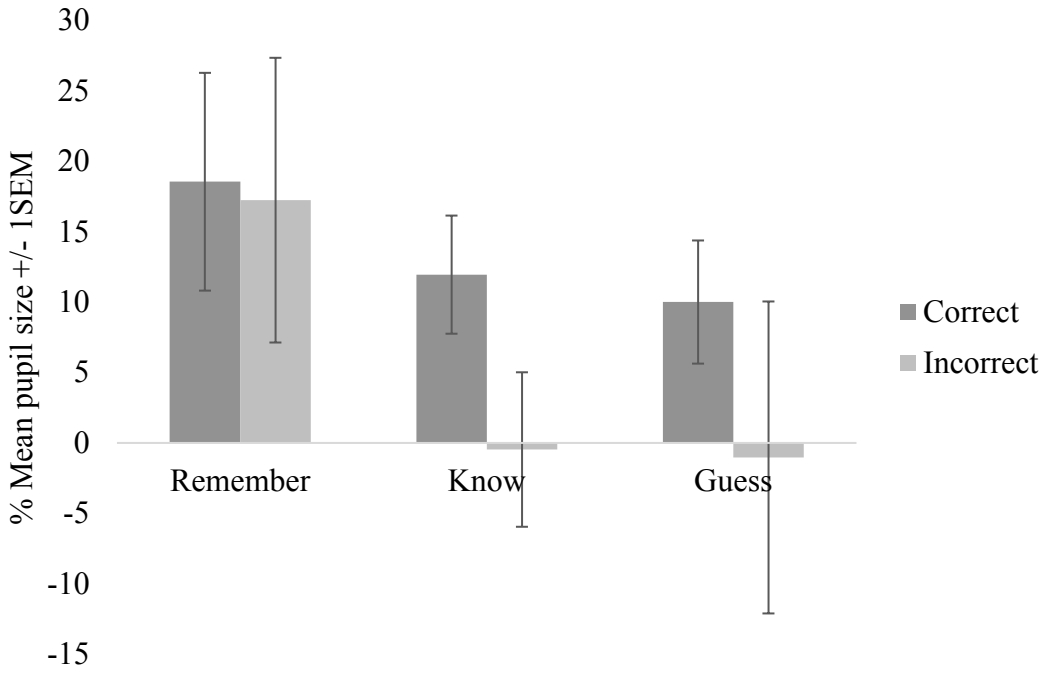


Fig. 56. Mean pupillary difference between the target and distractors in the first lineup presentation, as a function of identification accuracy (correct and incorrect), and RKG rating (remember, know and guess).

(Legend:) Positive: mean pupil size was larger when looking at the target than at the distractors. Negative: mean pupil size was smaller when looking at the target than at the distractors.

Inspection of fig. 56 reveals that pupils of "Remember" participants responded to the target irrespective of whether or not the participant made the correct answer, but this was not the case for "Know" or "Guess" participants. However, the samples of "Remember" and "Guess" participants were very small, and the error bars were extremely large, so only the data from "Know" participants were reliable.

For the second lineup presentation, there were no significant effects.

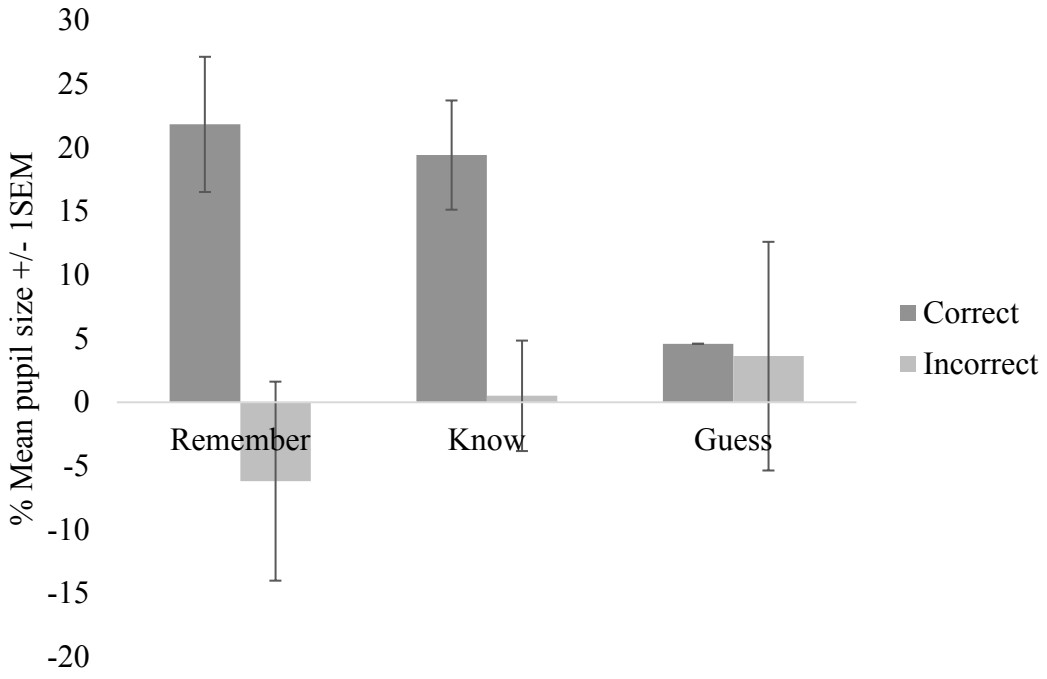


Fig. 57. Mean pupillary difference between the target and distractors in the second lineup presentation, as a function of identification accuracy (correct and incorrect), and RKG rating (remember, know and guess).

(Legend:) Positive: mean pupil size was larger when looking at the target than at the distractors. Negative: mean pupil size was smaller when looking at the target than at the distractors.

The combined data from both lineup presentations indicated that in "Know" participants, pupil sizes reflected identifications: pupils responded to the target by getting bigger in participants who identified him, but did not in those who failed to identify him. We did not have enough data to draw any conclusions from "Remember" or "Guess" participants in either presentation.

5.3.2. Target-Absent condition

Based on their identification response, participants were divided into "misidentifiers", (participants who mistook a distractor for the target), and “correct rejectors” (those who correctly responded that the target was absent from the lineup).

Three pupil size measures were taken from each participant. In participants who misidentified a distractor in the absence of a target, we wanted to measure pupillary responses to the face that was misidentified, so we treated this face as the "target", but called it the "false positive". Therefore, pupil sizes for distractors seen before the false positive were averaged together to produce a single mean pupil size measure, labelled “pre-false positive distractors (Pre)”. There was only one false positive, so only one measure was available for each participant, labelled “false positive”. Finally, pupil sizes for distractors seen after the false positive were averaged together to give a single mean pupil size measure labelled “post-false positive distractors (Post)”. In participants who correctly rejected all faces, we did not even have a false positive face to compare to the target. Therefore, we selected the distractor that had been misidentified most often (30% of the time) and designated this face to be the "false positive", as we considered it most likely to be considered a close match to the target and therefore most likely to elicit a large pupil size. Then, we followed the same procedure that we had followed for the target-absent misidentifiers.

To test whether pupil sizes changed in response to a misidentified face in the absence of the target, two two-way mixed ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with three levels: pre, false positive, and post) and one between-subjects factor, *identification response* (with two levels: correct rejectors and misidentifiers).

As can be seen in figs. 57 and 58. there were no significant effects in either presentation. Inspection of fig 57. indicated that pupillary responses might be different in misidentifiers and correct rejectors when looking at the target, but a one-way ANOVA confirmed that this was not the case, F (1,21) = 2.78, p = .11. Thus, pupillary responses did not discriminate between misidentifiers and correct rejectors or between face types in either presentation.

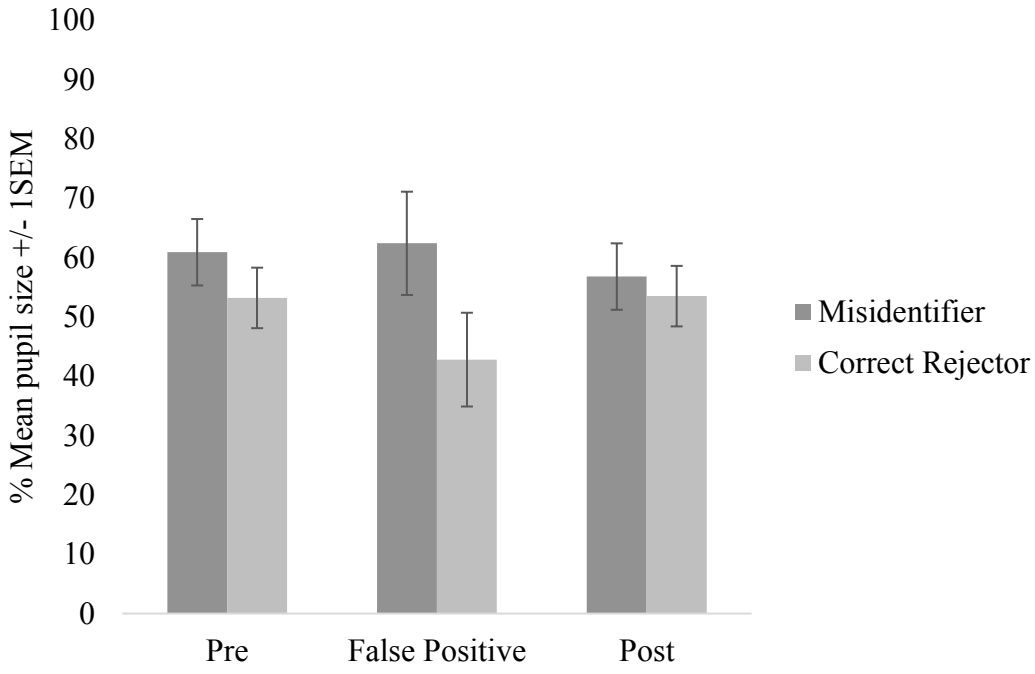


Fig. 58. Pupillary changes in response to the first lineup presentation:

(Legend:) Pupillary changes for pre-false positive distractors, false positive, and post-false positive distractors in the first lineup presentation, with participants grouped by identification response (correct rejectors and misidentifiers).

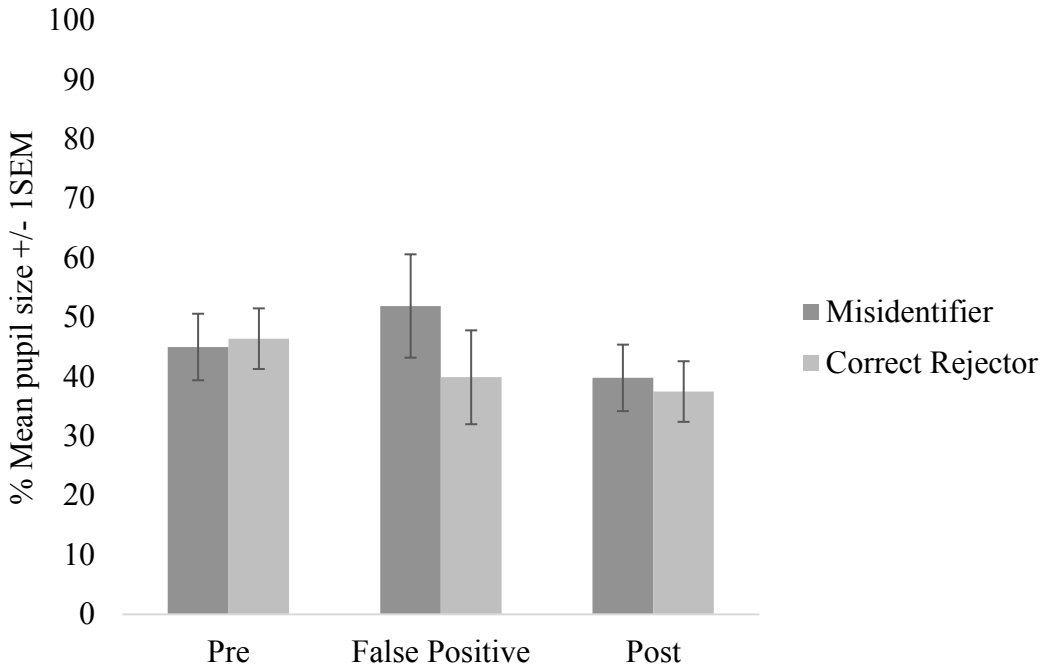


Fig. 59. Pupillary changes in response to the second lineup presentation:

(Legend:) Pupillary changes for pre-false positive distractors, false positive, and post-false positive distractors in the second lineup presentation, with participants grouped by identification response (correct rejectors and misidentifiers).

5.3.2.1. Using pupil size to predict identification response:

Two binary logistic regressions were used to determine whether pupil size change could predict whether participants made a correct or incorrect lineup decision. The predictor variable was pupil size (calculated as the mean difference between the target and the distractors) and the outcome variable was identification accuracy (correct or incorrect).

As expected, the logistic regression was not statistically significant for either lineup presentation: first lineup presentation, $\chi^2(1) = 2.44$, $p = .12$; second lineup

presentation, $\chi^2(1) = 3.14$, $p = .08$. Pupil size did not predict whether participants would correctly reject all the faces or misidentify a face.

5.3.2.2. Subjective assessments of identification accuracy:

Two Chi Square analyses were used to see whether participants' own assessment of their 'memory strength' was related to their actual lineup performance. There was no significant association between memory strength and performance for either lineup presentation: first lineup presentation, $\chi^2 (2) = 2.10$, $p = .35$; second lineup presentation, $\chi^2 (2) = 2.57$, $p = .28$. Therefore, participants' assessment of their memory was not a good indicator of their performance.

Table 7. Percentage of people (and raw frequencies) in each RKG response group to reject all the faces correctly or to misidentify a distractor, in the first lineup presentation.

	Correct	**Wrong**
Remember	70.00% (7)	30.00% (3)
Know	33.00% (2)	67.00% (4)
Guess	50.00% (3)	50.00% (3)

Table 8. Percentage of people (and raw frequencies) in each RKG response group to reject all the faces correctly or to misidentify a distractor, in the second lineup presentation.

	Correct	Wrong
Remember	50.00% (4)	50.00% (4)
Know	71.00% (5)	29.00% (2)
Guess	29.00% (2)	71.00% (5)

Next, we investigated whether pupillary changes were related to the RKG responses, taking into account whether or not the participant made a correct rejection. Two three-way ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with two levels: false-positive, and distractors) and two between-subjects factors, *identification accuracy* (with two levels: correct and incorrect), and *RKG rating* ("Remember", "Know" or "Guess"). The dependent variable was pupil size change.

There were no significant effects in either presentation. When taking into account whether a participant made a correct rejection, pupillary changes were not related to RKG responses in either lineup.

5.4. Discussion

The present research investigated whether pupillometry would benefit current UK police procedures, where eyewitnesses memorise the number of a face in a lineup if they wish to identify it. In Chapter 4, we found that pupillometry measured implicit

recognition in a hybrid lineup (with yes/no responses), so we anticipated that this would also be the case in the present research. Specifically, we predicted that participants who recognised the target in the target-present condition would have larger pupils when looking at his face compared to those of the distractors. We also predicted that there would be no significant pupillary changes in the target-absent condition. Both of our predictions were confirmed, but in contrast to the previous research, the pupillary changes in the target-present condition occurred with both lineup presentations.

An explanation for this effect may lie in competing accounts of pupillary responses to cognitive processing. This experiment made the assumption that pupillary changes were associated with memory strength, as pupils have been shown to be larger when memory strength is greater (Otero et al,, 2011; Papesh et al., 2012; Brocher & Graf, 2016) (see Goldinger & Papesh, 2012 for a review). This was the basis for using the RKG paradigm, as it is also based upon memory strength (e.g. Dunn, 2004; Wixted & Stretch, 2004; Dunn, 2008; Wixted & Mikes, 2010). In our previous research (Chapter 4), this assumption was supported by the data, as pupillary changes were only found for the first target-present lineup presentation (when *only* the target face had been seen before), and were greatest in participants who both claimed to remember his face *and* made an identification.

In this experiment, in the first lineup presentation, pupillary changes *only* occurred in participants who made an identification. Therefore, pupillary changes differentiated between participants who had identified the target and those who had not. In the second lineup presentation, pupillary changes differentiated between all three identification response groups: pupils again only responded to the target in participants who identified him; they gradually got slightly smaller in people who made no

identification; and they did not change in size in participants who misidentified a distractor. The main point is that in this experimental paradigm, both lineup presentations benefitted from the pupillary data.

However, why this occurred is not clear. In this second lineup presentation, all the faces would now have been familiar to the participants. Therefore, on the assumption that pupils respond to memory, the faces in the second lineup should all have elicited similar pupil sizes, but this was not the case. A memory strength account is nevertheless still plausible, as the target's face would have been the most familiar to those people who recognised him (as they had been exposed to his face more). Therefore, the pupil sizes of those who recognised him would have been somewhat larger when looking at his face compared with the less-familiar distractors. However, this account alone fails to explain why this did not occur in our previous experiment (Chapter 4).

Some clarity might lie in an account of cognitive load, to which pupils also respond: the greater the cognitive load, the larger the pupil size (Beatty, 1982; Jainta & Baccino, 2010; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; & see Ayres & Paas, 2012; Goldinger & Papesh, 2012; Murphy, Groeger, & Greene, 2016, for reviews). Perhaps the extra cognitive burden (albeit small) of having to memorise the number of the perpetrator might have produced a pupillary change in addition to the increase in pupillary size associated with recognising him. This small increase in pupil size associated with cognitive load might have been enough to make the pupil changes of those who recognised him in the second lineup statistically significant.

We also investigated whether pupil size change could predict whether participants made a correct or incorrect lineup decision, and found that pupil sizes successfully predicted participants' decisions in both lineup presentations of the target-present

condition. This is important, as finding a way of measuring an individual eyewitness's pupillary data, and comparing them to a model for predicting correct decisions could be useful to police, particularly when making assessments about the credibility of an eyewitness's identification response.

We were also interested in seeing whether participants' assessment of their 'memory strength' ("Remember", "Know" or "Guess") was related to their actual performance with the lineup, and found that it was (in both lineup presentations), but again only in the target-present condition. Again, this contrasted with our previous experiment (Chapter 4), where participants appeared to have no insight into the strength of their own memory. The most likely explanation for this difference is that participants were asked to respond in different ways between the experiments. It is probable that the Yes/No responses of the previous experiment constrained participants to retain a response choice that they were not always happy with. Indeed, this seemed to be the case. Although participants in the previous experiment were asked not to identify more than one face, many did make more than one identification, so their data were removed from analysis. The current UK police system successfully prevents people from identifying more than one person in the lineup, and allows participants to change their mind as the lineup progresses. Thus, participants were probably able to assess their performance better in this experiment.

Finally, we investigated whether pupillary changes were related to the RKG responses, taking into account whether or not the participant made a correct rejection. We found that they were, but only for the first lineup presentation of the target-present condition. However, we were unable to draw any firm conclusions, due to the sample sizes being too small in some groups.

Another thing we wanted to evaluate was the use of hybrid video identification systems in the UK. Previous research suggests that hybrid systems tend to increase the likelihood of misidentifying a perpetrator (Steblay et al., 2011). Therefore, although they are currently used, they are not recommended by researchers. Our previous research (Chapter 4) also showed that the second lineup presentation did not benefit from the pupillary data. However, in the present experiment, we found that the second lineup presentation was as useful as the first when it came to predicting whether a participant would identify the perpetrator or not. In fact, we found few differences between the presentations: participants tended to make the same identification response in both presentations, and most rated their memory strength the same on both occasions. However, the second presentation was no more useful, and considering the evidence that a second lineup presentation tends to increase errors (Steblay et al., 2011), it seems that the second lineup presentation is at best a waste of time.

Previous research suggests that the use of videos in UK police lineups is a good choice, particularly when using systems such as VIPER (n.d.). Wells et al. (2015), have shown that double-blind techniques, where the person administering the lineup is not aware of who is the suspect, have improved identification reliability, as this person is unable to influence the eyewitness. Software such as VIPER has made this even more effective as the eyewitness can proceed unassisted, by following instructions on the screen. VIPER makes lineups fairer, as it uses algorithms to select distractors on the basis of the physical appearance of the suspect. It is less prone to bias as a result (see Malpass, Tredoux, & McQuiston-Surrett, 2007 for a review of fair lineups). It is also beneficial as the distractors are taken from a database, and cannot be wrongfully convicted as a result (see Kemp, Pike, and Brace, 2001, for a commentary). Thus, the combination of VIPER and pupillometry could help to minimise the wrongful conviction of innocent people. The

use of video systems also has another advantage. The present research demonstrates that pupillometry has the potential to provide nuanced information about memory strength in lineups that is not possible with current methods. Although no research has yet been conducted to test this, it seems unlikely that pupillometry could be used with a simultaneous display for two reasons: first, because people might use peripheral vision to dismiss faces without even looking at them; and second, because it is unlikely that pupils can re-set between faces presented simultaneously, in the way that they can with sequential videos.

Therefore, it appeared that pupillometry could offer a supportive role in UK police procedures, as it can provide data that identification responses and RKG ratings fail to provide alone. Also, previous research has tended to find that using procedures that reduce the likelihood of misidentifying an innocent person means that the chances are increased that a guilty person goes free, or vice versa (Steblay, Dysart, Fulero, & Lindsay, 2001), but it seems that pupillometry could help to reduce both types of error. For instance, the pupillary data of some participants who failed to identify the target suggested that they might nevertheless have recognised him implicitly. One explanation is that implicit recognition did not reach the level of consciousness required to make an identification, another is that they did not want to identify the perpetrator. The data from participants like these could help reduce the non-conviction of guilty people. In contrast, some participants who did make an identification in the target-present lineup appeared to do so in the absence of implicit recognition, suggesting that they might have selected the target by chance, and this was supported by those who misidentified a distractor in the target-absent condition. Data from these participants could help to reduce the wrongful conviction of innocent people (who are suspected of a crime that they did not commit).

The measure we used needs to be refined, and more research is needed to find the most effective self-rating measure of memory strength, such as confidence ratings scales. Research into the application of pupillometry to US methods is also warranted. However, this research indicates that UK police procedures could benefit from introducing pupillometry to support measures already in place, as pupils appear to measure implicit recognition independently of established identification responses. Pupillometry is also unique as it can be applied simultaneously to the reduction of wrongful convictions and wrongful non-convictions, helping to reduce miscarriages of justice of both kinds.

References

Allport, G. W., & Postman, L. (1947). The psychology of rumor. New York: Russell & Russell.

Ayres, P., & Paas, F. (2012). Cognitive Load Theory: New directions and challenges. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364.

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology*, *53*(12), 1823–1835. https://doi.org/10.1111/psyp.12770

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291.

Conway, M. A., & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *The Quarterly Journal of Experimental Psychology Section A*, *48*(1), 125–140. https://doi.org/10.1080/14640749508401380

Cutler, B. L., & Penrod, S. D. (1988). Improving the reliability of eyewitness identification: Lineup construction and presentation. *Journal of Applied Psychology, 73*(2), 281-290.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-56.

Dunn, J. C. (2004). Remember-Know: A matter of confidence. *Psychological Review*, *111*(2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological review, 115*(2), 426-446.

Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: five days to execution and other dispatches from the wrongly convicted* (1st ed). New York: Doubleday.

Flowe, H., & Cottrell, G. W. (2011). An examination of simultaneous lineup identification decision processes using eye tracking. *Applied Cognitive Psychology*, *25*(3), 443–451. https://doi.org/10.1002/acp.1711

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Hagsand, A., Hjelmsäter, E. R. A., Granhag, P. A., Fahlke, C., & Söderpalm-Gordh, A. (2013). Bottled memories: On how alcohol affects eyewitness recall. *Scandinavian Journal of Psychology*, *54*(3), 188–195. https://doi.org/10.1111/sjop.12035

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Havard, C., & Memon, A. (2009). The influence of face age on identification from a video line-up: A comparison between older and younger adults. *Memory*, *17*(8), 847–859. https://doi.org/10.1080/09658210903277318

Havard, C., Memon, A., Laybourn, P., & Cunningham, C. (2012). Own-age bias in video lineups: a comparison between children and adults. *Psychology, Crime & Law*, *18*(10), 929–944. https://doi.org/10.1080/1068316X.2011.598156

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

The Innocence Project (n.d.), retrieved January 16th, 2018, from https://www.innocenceproject.org/cases/ronald-cotton/

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Kemp, R. I., Pike, G. E., & Brace, N. A. (2001). Video-based identification procedures: Combining best practice and practical requirements when designing identification systems. *Psychology, Public Policy, and Law*, *7*(4), 802–807. https://doi.org/10.1037//1076-8971.7.4.802

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, *44*(6), 894–904. https://doi.org/10.1111/j.1469-8986.2007.00566.x

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564.

Loftus, E. F., Schooler, J. W., Boone, S. M., & Kline, D. (1987). Time went by so slowly: Overestimation of event duration by males and females. *Applied Cognitive Psychology, 1,* 3–13. doi: 10.1002/acp.2350010103

MacLin, O. H., MacLin, M. K., & Malpass, R. S. (2001). Race, arousal, attention, exposure and delay: An examination of factors moderating face recognition. *Psychology, Public Policy, and Law*, *7*(1), 134–152. https://doi.org/10.1037//1076-8971.7.1.134

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 155–178). Lawrence Erlbaum Mahwah, NJ.

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783–792. https://doi.org/10.3758/BF03193074

Memon, A., Bartlett, J., Rose, R., & Gray, C. (2003). The aging eyewitness: Effects of age on face, delay, and source-memory ability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*(6), 338–345. https://doi.org/10.1093/geronb/58.6.P338

Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: the pupil's point of view. *Biological psychology*, *135*, 159-169.

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, *23*(5), 1316-1340. https://doi.org/10.3758/s13423-015-0982-5

National Academy of Sciences. (2014). Using eyewitness identifications: New report urges caution. *ScienceDaily*. Retrieved May 1, 2018 from www.sciencedaily.com/releases/2014/10/141002123735.htm

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory: Pupil size and recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Prehn, K., Heekeren, H. R., & van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Read, J. D. (1995). The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology*, *9*(2), 91-121.

Rush, E. B., Quas, J. A., Yim, I. S., Nikolayev, M., Clark, S. E., & Larson, R. P. (2014). Stress, interviewer support, and children's eyewitness identification accuracy. *Child Development*, *85*(3), 1292–1305. https://doi.org/10.1111/cdev.12177

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2009). The effect of retention interval on the Confidence–Accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*, *3*(9), 160300. https://doi.org/10.1098/rsos.160300

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78,* 22-33.

SR Research (n.d.), retrieved, April 23rd, 2018, from https://www.sr-research.com/products/eyelink-1000-plus/

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *25*(5), 459–473. https://doi.org/10.1023/A:1012888715007

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262–274. https://doi.org/10.1007/s10979-010-9236-2

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, *23*(2), 151–161. https://doi.org/10.1002/acp.1463

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE*, *7*(12), e51134. https://doi.org/10.1371/journal.pone.0051134

VIPER (n.d.), retrieved January 16th, 2018, from http://www.viper.police.uk

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*(6), 688.

Wells, G. L., & Olson, E. A. (2001). The other-race effect in eyewitness identification: What do we do about it? *Psychology, Public Policy, and Law*, *7*(1), 230–246. https://doi.org/10.1037//1076-8971.7.1.230

Wells, G. L., & Olson, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, *22*(6), 603.

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, *39*(1), 1–14. https://doi.org/10.1037/lhb0000096

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054. doi:10.1037/a0020874

Wixted, J. T., Read, D. J., & Lindsay, S. D. (2016). The Effect of Retention Interval on the Eyewitness Identification Confidence–Accuracy Relationship. *Journal of*

Applied Research in Memory and Cognition, *5*(2), 192–203. https://doi.org/10.1016/j.jarmac.2016.04.006

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11,* 616–641. doi:10.3758/BF03196616

Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: people are better with their own age. *Law and Human Behavior*, *26*(6), 641.

Wylie, L. E., Bergt, S., Haby, J., Brank, E. M., & Bornstein, B. H. (2015). Age and lineup type differences in the own-race bias. *Psychology, Crime & Law*, *21*(5), 490–506. https://doi.org/10.1080/1068316X.2014.989173

Yuille, J. C., & Tollestrup, P. A. (1990). Some effects of alcohol on eyewitness memory. *Journal of Applied Psychology*, *75*(3), 268-273.

Yuille, J. C., Tollestrup, P. A., Marxsen, D., Porter, S., & Herve, H. F. M. (1998). An Exploration on the Effects of Marijuana on Eyewitness Memory. *International Journal of Law and Psychiatry*, *21*(1), 117–128. https://doi.org/10.1016/S0160-2527(97)00027-7

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

CHAPTER 6. PIER PRESSURE: MEASURING IMPLICIT RECOGNITION IN FEARFUL EYEWITNESSES WITH PUPILLOMETRY

Abstract

Eyewitness misidentifications account for approximately 70% of wrongful convictions. Research has investigated why identification is error-prone, introduced procedures that reduce misidentifications, and measured credibility. However, behavioural responses can be contaminated by conscious decisions, and fail to index 'live' recognition of specific faces, while neurological measures are not practical. This study extends theoretical research, by using pupillometry to measure implicit recognition in a field study. Using a hybrid lineup procedure, we recorded pupillary responses with a portable eye-tracker as participants viewed the face of a researcher who they had just met, and distractor faces. Participants who identified her had larger pupil sizes when viewing her face than those who did not. We also manipulated anxiety level. Anxious participants had larger pupil sizes than non-anxious ones but both groups showed similar pupillary responses on the basis of recognition. The results suggest that pupillometry could be a practical tool for indexing individual recognition and predicting response accuracy in forensic settings.

Eyewitness identification is poor (the Innocence Project, n.d.; see also Dwyer, Neufeld, & Scheck, 2000), probably because eyewitnesses are required to recognise individuals often seen only briefly before, an extremely difficult task (Hancock, Bruce, & Burton, 2000). Variables that can affect eyewitness identification include "system" variables, procedural choices that can be modified to maximise the apprehension of perpetrators and which are amenable to improvement (see Wells et al, 1998; Wells & Olson, 2003 for reviews and recommendations) and "estimator" variables, over which the police have no control. Estimator variables include individual differences in eyewitnesses' ability, the specific characteristics of the event, and factors that affect the internal state of the eyewitness, such as alcohol and drugs (Yuille & Tollestrup, 1990; Yuille, Tollestrup, Marxsen, Porter, & Herve, 1998; Hagsand, Hjelmsäter, Granhag, Fahlke, & Söderpalm-Gordh, 2013), and stress (e.g. Valentine & Mesout, 2009; Rush et al., 2014).

Although it's a poorly-defined and rather nebulous concept, "arousal" has been suggested to show an inverted U-shaped relationship with performance, a phenomenon which has come to be known as the "Yerkes-Dodson law of arousal. Since then, researchers have investigated what gives rise to this effect, and how it can affect memory. For example, Easterbrook (1959) found that emotion narrows attention, so that relevant items are attended to, but irrelevant ones are not. This was tested in a series of experiments, such as those by Loftus & Burns (1982), who found that exposure to violence even affected recall of events immediately *preceding* the exposure. Later, Loftus, Loftus and Messo (1987) found that when presented with stimuli that contained a weapon, participants focused on the weapon, which affected subsequent recall and identification. This is a phenomenon known as ‘weapon focus’ (see Steblay 1992, for a review), where face recognition decreases in the presence of a weapon. However, it has been shown that a similar effect can occur with novel (unthreatening) items (Pickel,

1998). These findings lack ecological validity, given they are primarily based on the results of laboratory experiments using undergraduates viewing videos of staged crimes. Consequently they probably only tell us about unaffected *witness* memory rather than victim memory (Tollestrup, Turtle & Yuille, 1994). Deffenbacher, Bornstein, Penrod, and McGorty (2004) conclude that laboratory studies are likely to underestimate the negative effect of stress on eyewitness performance. So, other research has tried to investigate this in more naturalistic settings.

Peters (1988) found that people being immunised had a poorer memory for the face of the person wielding the hypodermic syringe than they did for the face of an aide who was not involved in the immunisation process, suggesting that someone is more likely to recognise people who were present at a fearful event than the person who was the source of fear. However, the effects of fear on recognition remain unclear. Yuille and Cutshall (1986) found that reported stress levels during a crime were not significantly related to subsequent recall, although they found that witnesses who had higher levels of stress were more exposed to the crime than those who experienced less stress, indicating *either* that stress enhances recall, *or* that their recall was affected by the amount of exposure to the crime. Finally, Valentine and Mesout (2009) tested participants in the London Dungeon, and found that only 17% of those who were anxious when they encountered a frightening actor identified him in a subsequent lineup, while 75% of those who were not anxious identified him correctly. These studies indicate the complexity of establishing the influence of anxiety on eyewitness identification.

Eyewitness research has also investigated ways to evaluate the accuracy of witnesses' identification performance. These have included the use of (separate) face recognition tasks, confidence ratings and neurological responses, but these have failed to

contribute significantly to accuracy prediction in forensic settings. For instance, Bindemann, Brown, Koyas, & Russ (2012) attempted to predict identification accuracy by comparing scores on established face recognition tests with lineup responses. They found that the 1-in-10 face recognition test (Bruce et al., 1999) provided a good index of eyewitness reliability for participants who made an identification (a correct identification or a misidentification), but not for those who made no identification (no identification or a correct rejection). They were therefore not useful in predicting people who subsequently missed a target (in a target-present display), or those who correctly concluded that the target was absent in a target-absent display. Moreover, face recognition scores can only provide the likelihood that an individual will be able to recognise faces in general; they cannot reveal anything about how likely it is that a specific face will be recognised.

Witness confidence has also been studied in an attempt to find a predictor of eyewitness performance. Confidence has some use when a witness makes a lineup identification (Sauerland & Sporer, 2009; Sauer, Brewer, Zweck, and Weber, 2009), as long as it is recorded immediately after the identification and before the witness receives any feedback about their performance. Very confident eyewitnesses also tend to have higher degrees of accuracy compared to unconfident witnesses (Brewer & Palmer, 2010). Not all research has found confidence to be a reliable guide to eyewitness accuracy (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988), as confidence ratings can be influenced by feedback, especially when recorded after a delay (Wells & Bradfield, 1998). However, recent research suggests that the confidence-accuracy relationship is more reliable than previously thought (Wixted, Read, & Lindsay, 2016).

The remember-know (RK) paradigm was created to measure states of awareness associated with memory retrieval, and has also been used in eyewitness research (e.g.

Sauerland & Sporer, 2009). It made the distinction between episodic memory ("Remember") and semantic memory ("Know") and has subsequently been refined by many researchers to also include "Guess" responses (Dunn, 2004). However, all three measures rely on behavioural responses and thus only measure explicit recognition, so they are both likely to be contaminated by conscious decision-making processes. The self-ratings scales are also likely to reflect self-esteem (Kröner & Biermann, 2007). As these identification responses are known to be unreliable, it is worth investigating whether responses that do not depend on conscious decisions are more reliable.

Therefore, researchers have investigated neurological responses, as these can measure implicit cognitive processes, and have been shown to be affected by a specific face (e.g. Lefebvre, Marchand, Smith, & Connolly, 2007). However, measuring neurological responses is not practical in applied settings, and inappropriate for testing recognition in victims of crimes who are stressed (see Miller & Bornstein, 2013).

One measure that may offer a practical and reliable indicator of eyewitness performance is pupillometry. Research has shown that pupils are larger when stimuli are emotional rather than neutral (e.g. Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008a; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016), larger when stimuli encourage goal-seeking (Mathôt, Siebold, Donk, & Vitu, 2015), when associated with reward (Satterthwaite et al., 2007), and larger when stimuli evoke stress, fear or are associated with trauma (Bitsios, Szabadi, & Bradshaw, 1996; Kimble, Fleming, Bandy, Kim, & Zambetti, 2010). The work of Goldinger, He, & Papesh (2009), and Goldinger & Papesh, (2012) suggests that pupillometry can also be used to measure cognitive fluctuations associated with face processing.

Also, pupil sizes appear to reflect memory strength. For instance, it has been found that pupils change size in the absence of an overt response (van Rijn, Dalenberg, Borst, & Sprenger, 2012), and can even occur despite efforts to deceive (Heaver & Hutton, 2011). Finally, we found that pupils responded to a target face in lineup (Chapters 4 & 5), and this was particularly striking when taking both participants' identification responses and their belief about their memory strength into account. The results showed that pupil size reflected memory strength and was independent of the explicit responses that participants gave.

However, the research described above was conducted in laboratories, so little is known about the use of pupillometry in forensic settings. Therefore, in the present study, we made four changes to the previous paradigm, taking inspiration from Valentine and Mesout (2009). First, we took the experiment out of the laboratory and recruited participants from the general public (at the British Science Festival, 2017). This was important, as victims and eyewitnesses stumble across crimes without expecting to have to recognise the perpetrators afterwards, so we tested people who happened across the experiment and were not expecting to identify someone that they had just met. Second, this field experiment required the use of an Eyelink Duo portable eye-tracker (SR Research, n.d.). In order to minimise anxiety, there is interest in conducting lineups in the homes of witnesses and victims, using virtual parade systems like VIPER (n.d.) (Miller & Bornstein, 2013). Testing the reliability of a portable eye-tracker that can assist with this is important. Third, we did not present participants with a mock crime scenario, but tested their ability to recognise the face of a researcher who conducted a questionnaire with them. This provided a to-be-recognised face in a natural scenario. Finally, we chose participants who had just ridden on the scary "Booster Ride" on Brighton Pier and separated them into two groups (anxious and non-anxious), based on their level of anxiety

during the ride, to see whether it affected responses. This was important, as victims and eyewitnesses can experience high levels of anxiety during a crime and during a lineup (Miller & Bornstein 2013).

We then tracked their pupils as they attempted to identify the researcher's face in a hybrid lineup procedure. We hypothesised that participants who correctly identified the target in a lineup would have larger pupils when looking at her face than at the faces of distractors. We also hypothesised that anxious participants would have larger pupil sizes than non-anxious participants, but did not know whether this would affect the ability to predict recognition in anxious participants. This study will provide further information on the effects of stress on eyewitness performance in a real-world setting, given that previous research has produced mixed results on this issue.

6.2. Methods

6.2.1. Participants

Twenty participants (12 males and 8 females) aged between eighteen and fifty-three (M = 27.35, SD = 8.14) were recruited from the Booster Ride on Brighton Pier during the British Science Festival (2017).

6.2.2. Apparatus and Materials

The stimuli consisted of 8 colour video clips of white female faces (aged approximately 20-30) that showed the head and shoulders against a white background. In each clip, the individual initially faced the camera. Then they turned their head slowly to the right, back to centre, to the left and back to centre. The video clips were constructed according to the criteria for lineup video clips set by the VIPER Unit of the West

Yorkshire Police (VIPER, n.d.). Videos were cropped and matched for size (17.5 x 13.3 cm), resolution (768 x 576 pixels), duration (10 seconds), and luminance. The lighting levels of the tent in which the study was performed were controlled.

Experiment Builder was run on a 21.5 inch laptop computer and an EyeLink Duo eye-tracker, which uses an infrared camera to provide precise measures of gaze location and pupil size. The head was stabilised during eye-tracking by means of a chin-rest, at an approximate distance of 60 cm from the computer screen that displayed the stimuli. The right eye was tracked for all participants.

A questionnaire consisting of 20 multiple choice questions was used to ascertain the emotional state of participants while they were on the ride, and one question asked if they had been anxious or not. The questionnaire took approximately two minutes to complete.

The Booster Ride is a scary ride on Brighton Pier that launches two pairs of people high into the air and swings them back down at high speeds while closely missing the ground. The ride lasts approximately five minutes.

The British Science Festival is a science festival that aims to connect people with scientists, engineers, technologists and social scientists (n.d.). In 2017, they were based in Brighton and had an evening event called the Brighton Pier Takeover, where different experiments, presentations and events were set up. We conducted our experiment at this event, taking advantage of the Booster Ride.

6.2.3. Design

This study used a mixed design: independent measures on identification response (with three levels: identifiers, non-identifiers, and misidentifiers) and anxiety level (with two levels: anxious and non-anxious) and repeated measures on face type (with three levels: pre-target faces, target face, and post-target faces). The dependent variables were pupil size, calculated as a percentage of each participant's overall pupil size range during the experiment (see 6.3. for details).

6.2.4. Procedure

People waiting to ride on the Booster Ride were approached by a female researcher for recruitment purposes. After the ride, those who had agreed to take part in the experiment were asked by her to complete a short questionnaire designed to evaluate their level of anxiety while they were on the ride. She spent approximately 2-3 minutes with each participant before they proceeded to the lineup, but was present in the vicinity of the ride throughout the event.

Following completion of the questionnaire, the participant was directed to a tent set up for the eye tracker, where the eye tracking element of the experiment was conducted by a second researcher. Participants did not see the first researcher again. At this point the participant placed their chin on the chin rest, and their eye movements were calibrated to five points on the computer screen. After reading instructions on the screen, their gaze was monitored with a drift check, which involved looking at a black dot on a white screen. They were told that they would be presented with a lineup that might include the face of the researcher who had just administered the questionnaire to them. They then saw a video lineup that included her face and seven other distractor faces.

The administration of the lineup followed current UK police procedures: the faces in the lineup were presented sequentially, but the lineup was shown twice. Each clip was assigned a number from 1-8 that was displayed clearly on the screen, and each clip played for 10 seconds. The ISI was not fixed, as each clip was separated by a drift check, which took approximately 2-3 seconds. Each participant saw the 8 lineup video clips in a different pseudo-random order (the target was never in the first two or last two faces in the lineup). After the final clip, participants were asked to provide their response: if they wanted to make an identification they had to type the number of the face that they thought was the researcher's, and if they did not want to make an identification they had to type '0'. After they had provided their response, they were asked to provide an RKG response for their performance (1 = remember, 2 = know, 3 = guess). Then the procedure was repeated, with the clips displayed in a different pseudo-random sequence, and participants were asked to make responses as they had for the first lineup presentation. They were told that this response could be the same as it had been in the first lineup presentation, or that they could make a different response if they wished. The eye-tracker recorded eye movement data and responses as participants viewed the clips.

After data collection, participants were grouped according to their level of anxiety. The basis for these groups was made from the final question in the questionnaire, which asked them to rate their level of anxiety while on the ride, on a scale from 1-5 (1 = not anxious at all, and 5 = very anxious).

6.3. Results

To standardise pupil size measurements between participants, the following procedure was used. The eye-tracker recorded a mean pupil size for each face. For each

participant, we converted this to a percentage of their pupil size change during the experiment, by identifying the face that elicited the largest mean pupil size and the face that elicited the smallest mean pupil size, and calculating the difference between them. The mean pupil size for each face was then calculated as a percentage of that difference. We did this to convert the arbitrary scores produced by the eye-tracker to a meaningful figure, and to standardise pupil size changes between participants.

From these values, three pupil size measures were produced for each participant. First, pupil sizes for distractors seen *before* the target were averaged together to produce a single mean pupil size measure, labelled “pre-target distractors”. There was only one target, so only one measure was available for each participant, labelled “target”. Finally, pupil sizes for distractors seen *after* the target were averaged together to give a single mean pupil size measure labelled “post-target distractors”.

In order to determine whether pupillometry could measure memory strength in a lineup in a field study with a portable eye-tracker, two two-way mixed ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with three levels: pre-target distractors, target, and post-target distractors) and one between-subjects factor, *identification response* (with three levels: identifiers, non-identifiers, and misidentifiers).

Participants were divided into three categories based on their identification response: "identifiers", participants who correctly identified the target; "non-identifiers", participants who mistakenly thought the target was absent; and "misidentifiers", participants who mistook a distractor for the target.

6.3.1. First lineup presentation, participants grouped by identification responses (identifiers, non-identifiers, and misidentifiers).

For the first lineup presentation, there was an effect of *identification response*, F (2, 17) = 4.22, p = .03, r = .45, η^2 = .33 (Identifiers: M = 65.38, SE = 4.87; Misidentifiers: M = 43.40, SE = 6.89; Non-identifiers: M = 47.83, SE = 6.89), but there was no significant effect of *face type*, F (2, 34) = 2.30, p = .11. There was also no significant interaction between *face type* and *identification response*, F (4, 34) = 2.53, p = .06.

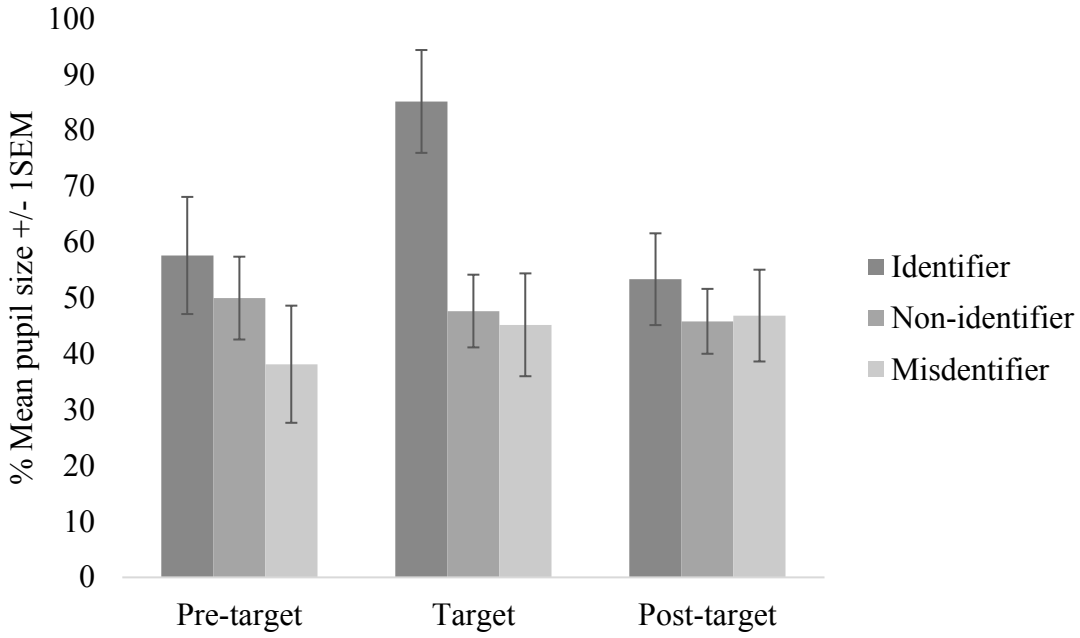


Fig. 60. Pupillary changes in response to the first lineup presentation:

(**Legend:**) Pupillary changes for pre-target distractors, target, and post-target distractors in the first lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

We conducted *t*-tests to clarify the results. As can be seen in fig. 60, only participants who correctly identified the target showed larger pupils when viewing her face than when viewing the distractors, t (9) = 4.38, p = .01 (distractor: M = 57.20, SE = 4.67; target: M = 85.18, SE = 5.10). Their pupils were also larger than those of the participants who did not identify her (identifier: M = 65.38, SE = 4.87; misidentifier: M = 43.40, SE = 6.89; non-identifier: M = 47.83, SE = 6.89).

6.3.2. Second lineup presentation, participants grouped by identification responses (identifiers, non-identifiers, and misidentifiers).

For the second lineup presentation, there were no significant effects: *face type*, F (2,34) = 2.52, p =.10 (Pre: M = 53.81, SE = 5.59; Target: M = 44.42, SE = 6.33; Post: M = 41.54, SE = 4.65); *identification response*, F (2,34) = 0.12, p = .89 (Identifiers: M = 46.83, SE = 5.12; Misidentifiers: M = 43.76, SE = 7.59; Non-identifiers: M = 49.19, SE = 8.48). There was also no significant interaction between *face type* and *identification response*, F (4,34) = 0.88, p = .49.

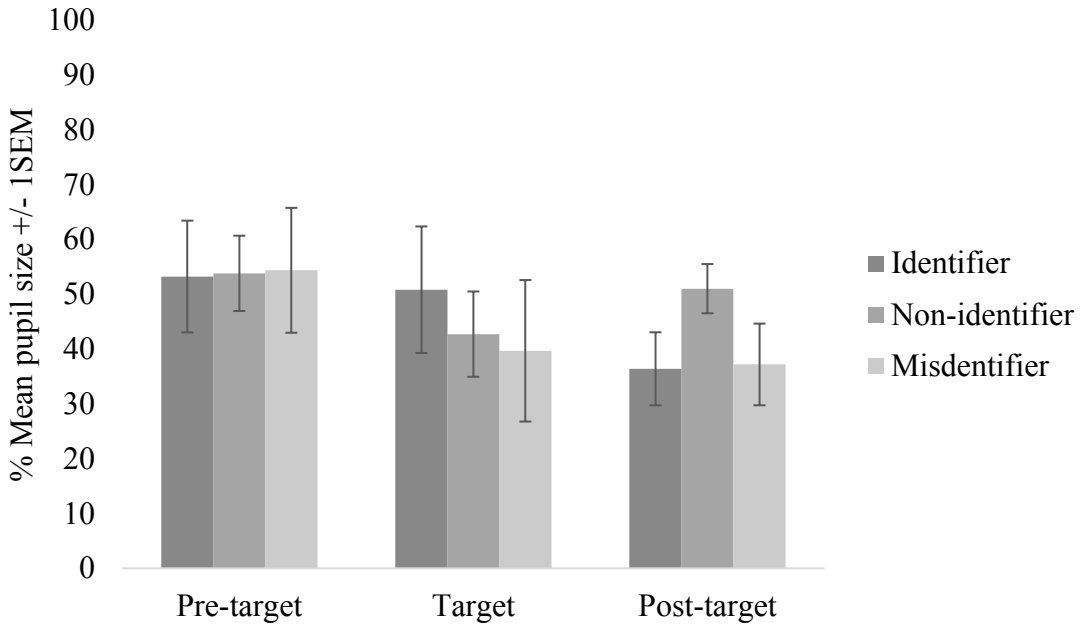


Fig. 61. Pupillary changes in response to the second lineup presentation:

(**Legend:**) Pupillary changes for pre-target distractors, target, and post-target distractors in the second lineup presentation, with participants grouped by identification response (identifiers, non-identifiers, and misidentifiers).

As can be seen in fig. 61, pupillary responses were no different between participants when they were divided according to their identification response, or between the different face types.

6.3.3. Using pupil size to predict identification response.

Two binary logistic regressions were used to determine whether pupil size change could predict whether participants made a correct or incorrect lineup decision. The predictor variable was *pupil size* (calculated as the mean difference between the target and the distractors) and the outcome variable was *identification accuracy* (correct or incorrect).

For the first lineup, the logistic regression model was statistically significant, $\chi2(1) = 8.50$ $p = .01$. The model explained 46.2% (Nagelkerke R^2) of the variance in lineup decision outcome (i.e. whether or not participants were correct in their decision) and correctly classified 75% of cases. For the first lineup presentation, pupil size was therefore a good measure of identification performance. This was not true for the second lineup, for which the logistic regression was not significant, $\chi2(1) = 3.15$, $p = .08$.

6.3.4. Participants' subjective assessments of identification accuracy:

Two Fisher's Exact analyses were used to see whether participants' assessment of their 'memory strength' (in terms of their "Remember", "Know" or "Guess" responses) was related to their actual performance with the lineup. There was no significant association between memory strength and performance in either lineup presentation: first lineup presentation, $\chi2\ (2) = 2.07$, $p = .58$; second lineup presentation, $\chi2\ (2) = 0.47$, $p = .43$. Therefore, participants' assessment of their memory was not a good indicator of their performance.

Next, we investigated whether pupillary changes related to the explicit (RKG) responses, taking into account whether or not the witness made a correct identification. Two three-way ANOVAs were used for analysis of pupil size measures in response to each lineup. For each, there was one within-subjects factor, *face type* (with two levels: target, and distractors) and two between-subjects factors, *identification accuracy* (with two levels: correct and incorrect), and *RKG rating* ("Remember", "Know" or "Guess"). The dependent variable was pupil size change.

For the first lineup presentation, there were no significant main effects of *face type,* $F\ (1,15) = 1.98$, $p = .18$. There was an interaction between *face type* and

identification accuracy, F (1,15) = 5.16, p = .04, η^2 = .26. However, there were no interactions between f*ace type* and *RKG*, F (2,15) = 0.28, p = .76, or between *face type*, *accuracy* and *RKG*, F (1, 15) = 0.11, p = .75. Univariate analysis also revealed a significant effect of *identification accuracy*, F (1,15) = 6.12, p = .03, r = .48: (correct: M = 68.14, SE = 5.57, incorrect: M = 44.36, SE = 7.84). There was no effect of RKG (p =.42).

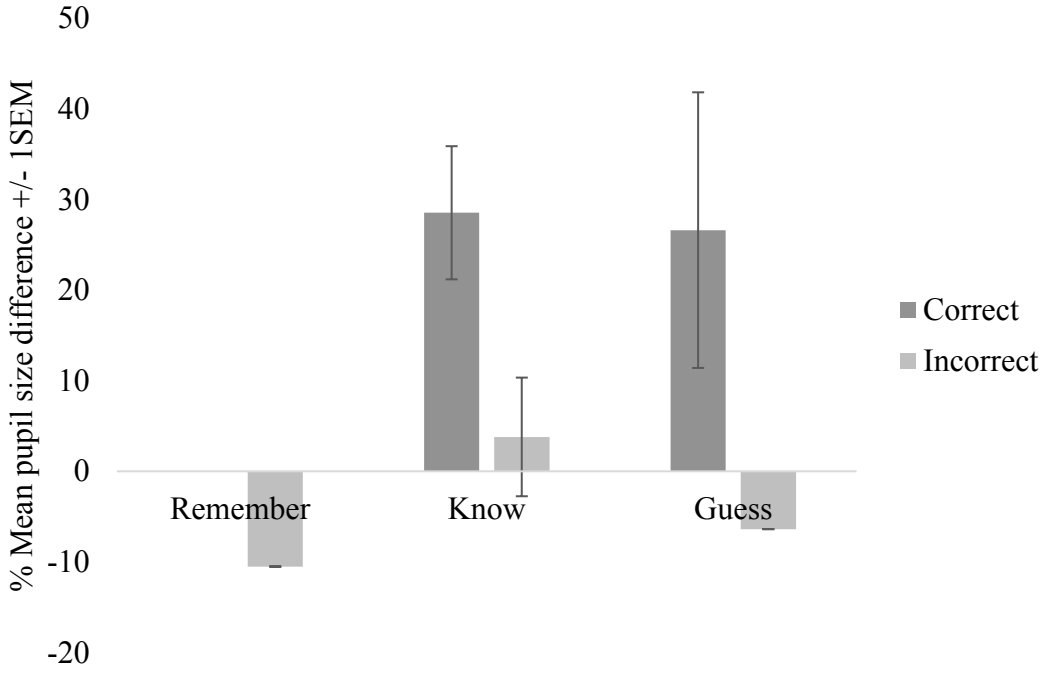


Fig. 61. Mean pupillary difference between the target and distractors in the first lineup presentation, as a function of identification accuracy (correct and incorrect), and RKG rating (remember, know and guess).

(Legend:) Positive: mean pupil size was larger when looking at the target than at the distractors. Negative: mean pupil size was smaller when looking at the target than at the distractors.

There was only one participant who responded "Remember" and they were incorrect, and only one who guessed and was incorrect. Therefore, the analysis was incomplete. However, as seen in fig. 69, it did tell us something about the "Know" responders, pupil size changes in response to the target were consistent with explicit identification decisions when the participant was correct, but not when the participant was incorrect. Pupil sizes were 29% larger when looking at the target compared to the distractors. However, in those who failed to identify the target despite saying that they knew him, pupil sizes were only 4% larger.

For the second lineup presentation, there no significant effects. As pupillary effects had been found in the first lineup presentation that were absent in the second, a final analysis was run to see whether identification performance was also different between the two presentations.

6.3.5. Identification accuracy for both lineup presentations.

A McNemar test was used to see whether participants' identification responses changed between the first and the second lineup presentations. There was no significant difference in performance between lineup presentations, $p = .50$ (2 sided): participants generally responded the same way in both lineup presentations.

Having established that pupils responded to the target, and predicted identification of the target overall, we then investigated whether there were differences in pupillary responses and accuracy in participants as a function of anxiety.

6.3.6. Pupil sizes in anxious and non-anxious participants.

6.3.6.1. Anxiety Groups

Due to adverse weather conditions, we only managed to recruit 20 participants. One participant failed to record their level of anxiety, so the responses of the other 19 participants were analysed. 12 of these rated their level of anxiety as 1 (not anxious at all), and 7 chose a rating that was greater than 1, indicating that they experienced at least some anxiety. Therefore, we decided to divide participants into two anxiety groups: "anxious" (7 participants) and "non-anxious" (12 participants).

A two-way mixed ANOVA was performed to see whether pupil sizes were different when people with different levels of anxiety looked at the faces in both lineup presentations. There was one within-subjects factor, *lineup presentation* (with two levels: first lineup presentation and second lineup presentation) and one between-subjects factor: *level of anxiety* (with two levels: anxious and non-anxious).

There was a significant effect of *anxiety* on pupil size, F (1, 17) = 8.97, p = .01 r = .59 (anxious: M = 57.86, SE = 3.69; non-anxious: M = 43.95, SE = 2.82): the pupil sizes of anxious participants were larger than those of non-anxious participants. There was no interaction between *lineup presentation* and *anxiety*, F (1, 17) = 0.01, p = .93.

6.3.7. Associations between pupillary changes and level of anxiety:

6.3.7.1. Pupillary changes

For all participants, regardless of anxiety level, we calculated the mean difference in pupil size between viewing the target and distractors. In the first lineup presentation when using all participants' data this difference was 15% (target: M = 66.80, SE = 6.22; distractors: M = 51.14, SE = 3.89), t (19) = 2.88, p = .01, r = .50. We therefore categorised all participants with a pupillary change of 15% or more when viewing the target as

"pupillary changers" and those with a pupillary change of less than 15% when viewing the target as "pupillary non-changers".

To see whether anxiety was associated with pupillary changes, Fisher's Exact tests were performed on level of anxiety (anxious and non-anxious) and pupillary changes (pupillary changers and pupillary non-changers). For the first lineup presentation, there was no significant difference in pupillary changes between groups, χ_2 (1) = 0.09, p = .57 (non-anxious: 50% = pupillary change, 50% = no pupillary change; anxious: 43% = pupillary change, 57% = no pupillary change). This analysis was not conducted on the second lineup data, as there was no difference in pupil size between the two face types (target and distractor).

6.3.8. Associations between identification accuracy and anxiety level:

To see whether identification accuracy was associated with anxiety level, Fisher's Exact tests were performed on level of anxiety (anxious and non-anxious) and accuracy responses (correct identification and incorrect identification) for the first lineup presentation. There was no significant difference in accuracy between anxiety groups, χ_2 (1) = 1.57, p = .22 (non-anxious: 42% correct, 58% incorrect; anxious: 71% correct, 28% incorrect). There was also no significant difference in accuracy between groups in the second lineup, χ_2 (1) = 0.83, p = .34 (non-anxious: 50% correct, 50% incorrect; anxious: 71% correct, 28% incorrect).

6.4. Discussion

We investigated whether pupil sizes could be used to measure implicit recognition in a lineup paradigm and found that they could, but only for the first lineup presentation:

participants who identified the target had larger pupils when viewing her face than when viewing the faces of distractors in the lineup. We also asked whether pupillometry could be used to predict accuracy in a lineup, and found that it could for the first lineup presentation: the model correctly predicted the response of 75% of participants.

We also investigated whether pupillometry would be an appropriate identification measure to use with anxious people, and found that it was in this experiment. While anxiety was associated with larger pupil sizes overall (Kimble et al., 2010), it made no difference to the fluctuations in pupillary responses associated with recognition. Finally, we wanted to test the use of a portable eye-tracker for forensic purposes that included anxious people, and found that it was ideal. The EyeLink Duo (SR Research, n.d.) was easy to use and set up, was able to process participants quickly, and provided precise pupillary data. It was also unobtrusive and allowed participants to move relatively freely during the experiment. It would therefore be a valid addition to police procedures, by providing them with additional pupillary data (to add weight to eyewitness credibility), and by being practical to use in the homes of witnesses and victims, reducing the anxiety associated with police lineups (Miller & Bornstein, 2013).

The pupillary results supported previous research, as the pupillary responses of accurate participants were different from those of inaccurate participants, and indicated that the former had implicitly recognised the target. They also helped to support the idea that the pupillary responses reflected implicit recognition rather than e.g. attraction. In our previous research (Chapters 4 and 5), the target was matched by VIPER (n.d.) on the basis of physical appearance, and the VIPER results were checked by trained police on the basis of age, race and e.g. attractiveness. By achieving similar pupillary results in the present experiment, using a different target and different distractors, we can be more

confident that recognition elicited the pupillary changes rather than attraction. However, there were issues with this experiment. Although the equipment allowed a fast turnover of participants, poor weather conditions on the day of the science event meant that we only managed to get 20 participants. As a result, we were not able to analyse RKG responses or levels of anxiety adequately. Both are discussed in turn below.

In this research, as described above, we found that pupillary responses predicted accuracy well. However, we were unable to make satisfactory use of the RKG responses to explore this much further. We did find that RKG responses were not related to participants' identification of the target. However, our results also showed that 'know' participants had large pupil size differences when they correctly identified the target, but not when they were incorrect, suggesting that explicit belief about memory strength (RKG) is related to pupillary changes, when taking accuracy into account. This supports our previous research (Chapters 4 and 5), which showed that when participants believe that they remember the target well (according to their RKG response), and do indeed remember the target (according to their identification response), their pupil sizes reflect their memory strength. However, when they *mistakenly* think they remember the target, their pupils do not change, despite their belief. The effect was similar (albeit smaller) in participants who 'knew' the target, but it was absent in participants who guessed, suggesting that they were indeed guessing.

When trying to analyse the effect of anxiety on identification and pupillary responses, we also encountered issues with the sample size, and we were unable to include anxiety as a variable in the main analyses. However, we were able to draw some conclusions based on the analyses that we did. Our results were consistent with previous research in showing that pupil sizes were larger in anxious participants (Kimble et al.,

2010), but this did not affect the pupillary fluctuations when viewing the faces, suggesting that the use of pupillometry in a lineup is viable with anxious participants. We also found no significant differences in accuracy between groups, conflicting with Valentine & Mesout (2009).

The current study extends the work of Valentine & Mesout (2009) in several ways. They tested recogniton of the face of the person who had been the source of the anxiety, and this is important, as eyewitness recognition often requires recognition of an obvious perpetrator. However, in our study, we investigated the effect of anxiety on recognising the face of someone who had not caused the anxiety. This is also important, as con-artists commit crimes without frightening their victims; and in some frightening crimes, it can be difficult to know who the perpetrator was. In Valentine and Mesout's study, anxiety negatively affected recognition of a frightening person, and Peters (1988) found that neutral faces were recognised *more* accurately than anxious ones, whereas in our study, the lack of power produced a non-significant result. Nevertheless, the source of the anxiety still remains an important factor in face recognition. However, there are other possible explanations for the differences between the results in our study compared to that of Valentine and Mesout.

For instance, while Valentine & Mesout (2009) tested heart rates to measure anxiety levels, we carried out short self-report questionnaires that may have been less reliable. Secondly, the length of time that participants experienced the anxious experience was different. In our study, they were on the ride for about five minutes, whereas the time taken to walk through the London Dungeon labyrinth would have been longer (see MacLin, MacLin, & Malpass, 2001, for a review). However, even the effects of exposure to a fearful event or person are not clear. For example, Yuille and Cutshall (1986) found

that stress was not negatively related to recall, as those who were closer to a crime or had longer exposure to it were likely to recall *more* items.

Third, we conducted the lineup immdediately after the questionnaire, so our participants only had 2-10 minutes between the Booster Ride and the lineup, whereas Valentine and Mesout's (2009) participants performed the lineup task after a considerably longer delay (at least 45 minutes). The effect of delay on memory is well-documented (MacLin et al., 2001), and may have been stronger in their anxious participants. Fourthly, by the time that Valentine and Mesout conducted their line up, participants were no longer anxious about the scary actor, but in our study the pupillary differences between the anxious and non-anxious participants indicated that the anxious participants were still anxious (Kimble et al., 2010). Thus, Valentine and Mesout's study reflects the procedual issues of eyewitness identifcations better than this study in terms of delay, as eyewitnesses sometimes have to wait months or even years to do a lineup. However, our study reflects the effects of lineup anxiety more effectively. It has been shown that doing a lineup can be extremely stressful for eyewitnesses, as there is pressure to help the police make a conviction, yet there is concern about the consequences of wrongful convition. Also, if a traumatised victim sees the perpetrator, this is also highly stressful (Miller & Bornstein, 2013).

Finally, while Valentine and Mesout's (2009) study focused on the effects of anxiety on identification, this study focused on the use of pupillary responses to predict identification accuracy, and whether this was affected by anxiety. Therefore, these studies combine to shed light on the complexities of eyewitness identification and the usefulness of naturalistic studies.

In short, pupillary responses provide a practical solution to predicting eyewitness identification accuracy with lineups, that can help to determine credibility. This is because they measure implicit recognition processes as participants look at specific faces. Pupillometry also appears to be valid and appropriate for anxious eyewitnesses. Pupillary responses to the individual faces in the lineup do not seem to be affected by anxiety, and the portable equipment is unobtrusive and suitable for lineup procedures outside the police station, reducing the anxiety of eyewitnesses as they view the lineup. This pilot study suggests that there is promise in using pupillometry in forensic settings, and provides a positive first step to more research in this area.

References

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Bitsios, P., Szabadi, E., & Bradshaw, C. M. (1996). The inhibition of the pupillary light reflex by the threat of an electric shock: a potential laboratory model of human anxiety. *Journal of Psychopharmacology*, *10*(4), 279–287.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology*, *15*(1), 77–96. https://doi.org/10.1348/135532509X414765

British Science Festival (2017), retrieved 8th February 2018, from https://www.britishscienceassociation.org/british-science-festival

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339–360.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-56.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological review*, *66*(3), 183-200.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*(6), 687-706.

Dunn, J. C. (2004). Remember-Know: A matter of confidence. *Psychological Review*, *111*(2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524

Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: five days to execution and other dispatches from the wrongly convicted* (1st ed). New York: Doubleday.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105–1122. https://doi.org/10.1037/a0016548

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Hagsand, A., Hjelmsäter, E. R. A., Granhag, P. A., Fahlke, C., & Söderpalm-Gordh, A. (2013). Bottled memories: On how alcohol affects eyewitness recall. *Scandinavian Journal of Psychology*, *54*(3), 188–195. https://doi.org/10.1111/sjop.12035

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

The Innocence Project (n.d.), retrieved 18th May, 2018, from https://www.innocenceproject.org/causes/eyewitness-misidentification/

Kimble, M. O., Fleming, K., Bandy, C., Kim, J., & Zambetti, A. (2010). Eye tracking and visual attention to threatening stimuli in veterans of the Iraq war. *Journal of Anxiety Disorders*, *24*(3), 293–299. https://doi.org/10.1016/j.janxdis.2009.12.006

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept—Towards a model of response confidence. *Intelligence*, *35*(6), 580-590.

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, *44*(6), 894–904. https://doi.org/10.1111/j.1469-8986.2007.00566.x

Loftus, E. F., & Burns, T. E. (1982). Mental shock can produce retrograde amnesia. *Memory & Cognition*, *10*(4), 318–323.

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus.". *Law and Human Behavior*, *11*(1), 55-62.

Mathôt, S., Siebold, A., Donk, M., & Vitu, F. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, *144*(3), 513–521. https://doi.org/10.1037/a0039168

Miller, M. K., & Bornstein, B. H. (Eds.). (2013). *Stress, Trauma, and Wellbeing in the Legal System*. Oxford, UK: Oxford University Press. DOI: 10.1093/acprof:oso/9780199829996.001.0001

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Peters, D. P. (1988). Eyewitness memory and arousal in a natural setting. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory:*

Current research and issues: Vol. 1. Memory in everyday life (pp. 89–94). Chichester, UK: Wiley.

Pickel, K. L. (1998). Unusualness and threat as possible causes of "weapon focus". *Memory*, *6*(3), 277-295.

Prehn, K., Heekeren, H. R., & van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Rush, E. B., Quas, J. A., Yim, I. S., Nikolayev, M., Clark, S. E., & Larson, R. P. (2014). Stress, interviewer support, and children's eyewitness identification accuracy. *Child Development*, *85*(3), 1292–1305. https://doi.org/10.1111/cdev.12177

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage*, *37*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2009). The effect of retention interval on the Confidence–Accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

SR Research (n.d.) retrieved February 8th, 2018, from http://www.sr-research.com/eyelinkportableduo.html

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior*, *16*(4), 413-424.

Tollestrup, P., Turtle, J., & Yuille, J. C. (1994). Eyewitness characteristics and memory: A survey of actual police cases. *Adult eyewitness testimony: Current trends and development.* New York: Springer-Verlag.

VIPER (n.d.) retrieved February 8th, 2018, from http://www.viper.police.uk

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, *23*(2), 151–161. https://doi.org/10.1002/acp.1463

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE*, *7*(12), e51134. https://doi.org/10.1371/journal.pone.0051134

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye:

Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360-376.

Wells, G. L., & Olson, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, *22*(6), 603-647

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *5*(2), 192-203

Yuille, J. C., & Cutshall, J. L. (1986). A case study of eyewitness memory of a crime. *Journal of applied psychology*, *71*(2), 291-301.

Yuille, J. C., & Tollestrup, P. A. (1990). Some effects of alcohol on eyewitness memory. *Journal of Applied Psychology*, *75*(3), 268-273.

Yuille, J. C., Tollestrup, P. A., Marxsen, D., Porter, S., & Herve, H. F. M. (1998). An exploration on the effects of marijuana on eyewitness memory. *International Journal of Law and Psychiatry*, *21*(1), 117–128. https://doi.org/10.1016/S0160-2527(97)00027-7

CHAPTER 7. THE EYES HAVE IT: DISCUSSING THE PUPILLARY EFFECTS OF FAMILIAR AND UNFAMILIAR FACE PROCESSING

7.1. Overview

This thesis had one main aim: to see whether pupillometry was useful in measuring cognitive fluctuations when processing faces with different degrees of familiarity. The results of the experiments consistently showed that pupillometry measured these fluctuations well. Even in Chapter 3, that had inconsistent results, there were some pupillary effects where other measures had none, suggesting that pupillary responses were more sensitive at detecting cognitive fluctuations than the other measures. The success in using pupillometry was most noticeable in the second half of the thesis, where forensic applications were explored, as pupil size changes were more clear-cut. These pupillary responses were attributed to recognition of the target in participants who remembered him, and this was most common in the first lineup presentation, when his face was the only familiar one in the lineup. Thus, we concluded that the pupillary changes were responding to a sense of familiarity.

The thesis also assessed the viability of using pupillometry in the field of face recognition in two separate domains: theoretical and applied. In terms of theories, it asked whether pupillary responses when looking at faces measured cognitive load, cognitive engagement or memory strength, and which theoretical construct best accounted for them. We found that all three provided accounts of the pupillary responses in different

ways, but that memory strength best accounted for the pupillary responses to the target's face in a lineup paradigm.

In terms of applications, we concluded that pupillometry has the potential to provide a practical tool for police when eyewitnesses view lineups of possible suspects. This is for three reasons. The first is because it measures responses that are independent of overt identification responses, which are known to be unreliable (Heaver & Hutton, 2011). The second is because it measures cognitive responses at the same time as the eyewitness is looking at a face, rather than after seeing all the faces. The third is because it measures the cognitive response to each face separately, rather than either providing the likelihood that a face will be recognised because of general face recognition ability (Bindemann, Brown, Koyas, & Russ, 2012), or has been recognised because of confidence in performance (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988), neither of which can be relied upon.

7.1.1. Theories

The thesis proposed three theoretical constructs to account for the pupillary changes: cognitive load, cognitive engagement, and memory strength. We theorised that face processing is affected by the familiarity of the face: either by producing variations in processing demands (reflected in variations in cognitive load); by the fact that faces of differing familiarity produce differences in the level of cognitive engagement; or by the fact that memory demands are less for familiar faces than unfamiliar ones. Given that pupil size is affected by all three processes, it seemed reasonable that facial familiarity might also affect pupil sizes as a result of these processes.

As the main construct of CLT, cognitive load (which is the notion that task demands affect mental workload) was initially devised to account for how the design of instructions affects cognitive load and learning (Sweller, 2010). Influences on cognitive load include the task demands (difficult tasks will have higher loads than easier ones) and things like poor instructions that increase cognitive load unnecessarily (see Moreno & Park, 2010; Sweller, 2010; Ayres & Paas, 2012; Murphy, Groeger, & Greene, 2016 for reviews). It has been shown that cognitive load can be measured with pupillometry (Jainta & Baccino, 2010; Piquado, Isaacowitz, & Wingfield, 2010; Chen & Epps, 2014; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014). The concept of cognitive load probably applies well to face processing, as processing faces that have only been seen briefly before is likely to place a greater cognitive burden on limited working memory resources than processing familiar faces does, so pupil sizes should be smaller when looking at familiar faces.

The second theoretical construct was cognitive engagement. This term is less well defined, but is based on the premise that salient objects (that have meaning to the observer, such as emotional content or social importance) will be more engaging than non-salient objects. There is evidence that pupil sizes are larger when looking at salient objects than non-salient objects (e.g. Partala & Surakka, 2003; Laeng & Falkenberg, 2007; Bradley et al. 2008; Võ et al., 2008; Prehn, Heekeren, & van der Meer, 2011; Snowden et al., 2016). Given that faces are socially-important, it is likely that they will also differ in terms of saliency. A socially-important face should be more salient than an unimportant one. Therefore, pupils should be larger when looking at a socially-important face than a socially-unimportant one.

Finally, we considered memory strength. This suggests that memory strength is determined by the similarity between the item held in the memory and the cue or target item, and that memory-based decisions are based on a strength of evidence continuum (Dunn, 2008). It has been shown that pupils are larger when memory strength is greater (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012). Therefore, faces that are more robustly represented (Hancock, Bruce, & Burton, 2000; Burton, Jenkins, & Schweinberger, 2011) should elicit a larger pupil size than weakly-represented (or novel) faces when they are recognised. This is because their mental representations contain more flexible information that can be matched to the target or cue.

In Chapter 2, the faces contained minimal emotional or salient content. Decline in cognitive load over successive trials provided the most plausible account of face learning, indicating that face learning was gradual. However, the reduction in pupil sizes could also be accounted for by changes in cognitive engagement, indicating that participants became increasingly bored by the task. The pupillary changes cannot be accounted for in terms of memory strength because the pupil sizes became *smaller* as the faces were familiarised, and an account in terms of memory would make the opposite prediction (Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Brocher & Graf, 2016; Goldinger & Papesh, 2012)..

In Chapter 3, we used unfamiliar, personally-familiar and own face images. Larger pupils were found as participants looked at their own face, indicating either that it engaged them the most (supporting accounts of own face biases, e.g. Kircher et al., 2000; Tacikowski & Nowicka, 2010), or that memory for one's own face is strongest. Both accounts are plausible. However, as the faces were not mirror-reversed, they could have

been more difficult to process than the other familiar faces, so we could not rule out an account in terms of cognitive load (Brédart, 2003).

By this point in the thesis, while we were unable to be sure about the relative contributions of cognitive load, cognitive engagement or memory strength to pupillary changes, we were confident that pupillary responses were good measures of the cognitive fluctuations involved in processing different face types, so we focused the second half of the thesis on exploring the practical application of pupillometry to forensic settings. However, we continued to evaluate the findings in light of the theoretical constructs.

In Chapter 4, we started a series of experiments investigating the viability of pupil size as a measure of implicit recognition in lineups. We also measured participants' beliefs about their memory strength, using the remember-know-guess (RKG) paradigm, so that we could compare participants' beliefs about their memory with a pupillary measure of their memory strength. In the experiments in Chapters 4-6, the evidence shifted fairly convincingly in the direction of memory strength: people who identified the target had much larger pupil sizes when looking at his face than when looking at other faces, suggesting that they remembered his face (although this could also have been because it engaged them more than the distractors' faces did). More support for an interpretation in terms of memory strength came from assessing participant's beliefs about their memory. We weighed their belief in their memory strength against the pupillary evidence, when taking into account participants' identification of the target. This showed (in Chapter 4) that when participants claimed to remember the target and were correct, their pupil sizes reflected their belief in this strong memory. However, when they *thought* they remembered the target and were *wrong*, their pupils did not respond to the target, indicating that their belief in their memory was wrong. This suggested that pupillary

measures of memory strength were more accurate than participant's belief in their memory. There were also no pupillary changes in the target-absent condition, also supporting an account in terms of memory. This is because when the target was not present there were no faces in the lineup that *could* have been remembered (at least in the first lineup presentation).

Therefore, while no definitive conclusion could be made about the theoretical accounts of pupillary responses, the evidence from this thesis suggests that cognitive load accounts for the difficulty in processing faces that contain no salient or emotional content, which decreases as the faces become more familiar, but when faces are familiar (and salient) pupil sizes fairly convincingly reflect memory strength (although an account in terms of cognitive engagement was also plausible to some extent).

Table 10. Pupillary evidence in support of the three theoretical constructs: Cognitive load, Cognitive engagement, and Memory strength.

Chapter	Large pupils	Small pupils	Cognitive Load	Cognitive Engagement	Memory strength	Explanation
2	Early trials	Later trials	✓	✓	✗	Pupils get smaller as faces become easier to recognise Pupils get smaller as participants lose interest in the task
	Unfamiliar Faces	Familiar Faces	✓	✗	✗	Unfamiliar faces are harder to process
	Asian Faces	Caucasian Faces	✓	✓	✗	Asian faces were harder to process. Asian faces may have been more engaging
3	Own faces	Familiar faces	✗	✓	✓	Own faces are more engaging than other faces. Own faces should be more robustly represented
4,5,6	Target face	Distractors	✗	✓	✓	Only the target face could have been remembered The target face was more engaging than the distractors, particularly if motivated to identify it.

There may be at least two possible reasons for the inconclusive data. It may be that these constructs are not mutually exclusive. For instance, when retrieving items that are robustly represented (strong memory), more mental effort (cognitive load) may be

required (if the mechanism through which cognitive load is increased or decreased in this context is via the robustness of the representation of the face being recognised). An alternative explanation is that pupillometry is not satisfactory for separating out their contributions. Thus, when strong memories and mental effort co-occur, pupillary effects are amplified, but if they compete, pupillary effects may be attenuated. Nevertheless, with further research into the phenomena, pupillometry appears to be a measure that has potential to measure the fluctuations in cognitive processing involved in face recognition.

7.1.2. Applications

This thesis was mainly concerned with forensic applications, but the results of the experiments in Chapter 2 also suggested that there may be some use for pupillometry in learning settings. It may be possible to use pupil size changes to assess teaching instructions and learning outcomes in terms of cognitive load. It may also be possible to assess the contribution of cognitive load on face learning in people with prosopagnosia. However, the practical aspect of this thesis was more concerned with the forensic applications of pupillometry.

Identification responses have been shown to be poor measures of eyewitness recognition, and are a major factor in miscarriages of justice (the Innocence Project, n.d.; see also Dwyer, Neufeld, & Scheck, 2000; Wells & Olson, 2003). It has also been shown that confidence ratings are not reliable measures of credibility (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988) although more recent research suggests that the confidence-accuracy relationship is stronger than previously thought (Wixted, Read, & Lindsay, 2016). Scores on recognition tests are also unreliable as they only show general face recognition ability and are ineffective when people make no identification (Bindemann et al., 2012). Also, all three measures (identification responses, face

recognition tasks and confidence ratings) are behavioural responses that cannot measure implicit recognition that conflicts with them (Heaver & Hutton, 2011). These might occur either because traces of recognition are not strong enough for the participant to be aware of them, or because they choose to ignore them (e.g. lying or low confidence). Pupillometry offers a practical solution to measuring these implicit responses when a face is recognised.

Having established that pupils respond to the implicit cognitive processes involved in face processing, we considered that pupillary responses might provide an additional tool to measuring face recognition in eyewitnesses while viewing individual faces in lineups, as they could provide information about the strength of recognition of a specific face that the other measures fail to provide. We conducted a series of experiments to test this, and found that the pupil sizes of participants who recognised the target responded to the target's face in each one. Pupil sizes also shed light on the efficacy of the second lineup presentation in standard police procedures.

Chapter 4 summarised our first attempt to test pupillary responses in a lineup, something that had not been done before (to our knowledge). In the target-present condition, pupil sizes were good predictors of identification. We also found that pupillary responses were good measures of memory strength of the target: in "Remember" or "Know" participants, pupillary responses reflected decision responses and indicated whether their belief in their memory was correct or not. However, in some participants who guessed but made no identification, there were pupillary changes to the target's face, indicating that they had recognised him implicitly, although they failed to identify him. Thus, pupillary responses were independent of explicit identification responses. Also, there were no pupillary effects in the target-absent condition, even in participants who

misidentified a distractor (as they were viewing the face that they misidentified), indicating that the pupillary changes seen in the target-present condition occurred because of the presence of the target.

In Chapter 5, we attempted to apply our experimental paradigm to UK police lineup procedures, to assess pupillometry in established forensic settings. Again, we found that pupil sizes were good predictors of identification of the target. They were again largest in participants who identified the perpetrator when looking at his face. However, in this case, the main effects were present in both lineup presentations. As this experiment was the first of its kind, and the findings run counter to psychological understanding about the purpose of the second lineup presentation in hybrid systems, it warrants further investigation. Previous research has shown that there are no advantages to having more than one lineup presentation, as any small improvements in accuracy are negligible and come at the expense of also producing more misidentifications (Steblay, Dietrich, Ryan, Raczynski & James, 2011). It is therefore important to know whether a pupillometry lineup that would be introduced to assist with assessing credibility would be most effective with just one lineup presentation or two.

The discrepancy between this experiment (Chapter 5) and that of Chapter 4 highlights the importance of following established procedures in experiments (and testing them against other procedures), but it did not change the most important finding: that pupillary responses measured implicit recognition of the target in a lineup, an effect that was absent when the target was not in the lineup.

In Chapter 6, we conducted the police-style lineup outside the laboratory, using a portable eye-tracker, an EyeLink Duo (SR Research, n.d.), and tested whether it was possible to test memory strength using pupillometry outside the lab. We also tested the

effects of anxiety on identification accuracy and pupillary responses, as understanding the influence of anxiety on eyewitness performance is important. We found that anxious participants had larger pupil sizes overall, but this did not compromise the effect of memory strength on them (pupil size *changes* were the same between anxiety groups). We had the same results in this experiment as in all the preceding experiments: the pupil sizes were largest in participants who identified the target when looking at her face, and pupillary responses reflected identification responses. As in Chapter 4, these effects were only found in the first lineup presentation. This adds weight to the argument that the second lineup presentation is unnecessary in police lineups. This is supported by the fact that there was no difference in accuracy between lineup presentations. This experiment indicated that pupillometry can be used forensically using a portable device, and is effective to use on anxious participants. This has positive implications for the application of pupillometry in this field, as it offers an ethical solution to gathering evidence. This is because anxious eyewitnesses and victims can view the lineup in their own home, rather than being stressed by doing so at the police station. The responsibility placed on eyewitnesses is known to be a burden, so any means that can reduce the negative effects is worth considering (see Miller & Bornstein, 2013, for a review). However, we did not get the number of participants that we were hoping for in this experiment, so this line of research also needs further investigation.

Overall, we suggest that pupillometry is a promising new approach to the issues with eyewitness identifications, as it can provide a practical additional tool to help police, lawyers and jury members make better-informed credibility assessments. This is because they provide nuanced information about the memory strength of the eyewitnesses as they are looking at a specific face, that can be independent of unreliable identification responses.

7.1.3. Limitations

In Chapter 2, the main issue was that we tested too many variables in a series of complex experiments. The aim was to conduct a broad investigation into whether pupillometry measured the process of face learning, and whether this clarified how faces are learnt and what influences it. Our main goal was to see whether pupillometry could detect differences between familiar and unfamiliar face processing. We had wanted to test the variables in separate experiments, but due to the difficulty in getting and keeping participants to test multiple times, a decision was made to test all participants in all conditions. This made it difficult to know where to separate one variable from another.

These experiments were also probably too easy. While a few participants performed poorly (hardly improving at all), most performed at or close to ceiling throughout, making it difficult to make assertions about improvements in accuracy. Therefore, it would have been better to have designed a more difficult experiment, and tested participants' face processing abilities beforehand. We could then have separated them into super-recognisers, typical processors, and prosopagnosics, to gain more insight into the learning patterns in people with different face recognition abilities.

Overall, the experiments were successful in terms of providing pupillary data that were subtler than the decision responses data, but teasing apart the multiple effects and interpreting the interactions was challenging. In the third experiment, the variables of age and race were removed, simplifying the process. However, an even simpler version could be attempted, by only using faces and participants of the same gender.

On balance, the experiments were useful, as by testing participants of a different age, we clarified the issues with using blink and pupillary data in older participants, as

the physical effects of aging made the data difficult to interpret. This was justified in the second experiment, where the effects that were probably related to physical aging in the first experiment were replaced by those that could be attributed to an asymmetrical Other Race Effect. It also provided us with a clearer idea of the direction we wanted to take with the subsequent experiments.

In Chapter 3, due to inexperience, the data were initially collected in slightly different ways at both universities, meaning that we had to re-assess data collection strategies after a considerable number of participants had already been tested. This was made more problematic due to the small departments at the universities and the annual turnover of the PhD students who constituted a large proportion of the participants. This meant that we were unable to test as many people as we wanted once we had established our new procedure. However, the experience was useful, as it demonstrated the importance of piloting experiments and clear instructions when collaborating.

The most problematic issue was the way that we presented the own face images. We chose not to mirror-reverse them, as like the other faces, they were taken (in original format) from Facebook or university profiles. While we speculated that people are now used to seeing their faces veridically as well as mirror-reversed nowadays, due to camera phones and social media, previous research suggests that people are more familiar with their own face in a mirror-reversed form than in its veridical orientation (Brédart, 2003). As the participants that we used were young and used social media often, we made the assumption that their own faces would be the most familiar to participants and hence the easiest to process. However, this assumption could have affected the findings somewhat. Even with the small sample, we may have achieved clearer results if the own face images had been mirror-reversed in this experiment.

However, despite the issues with this experiment, we still managed to detect pupillary and fixation differences between the face types, suggesting that participants' own faces were perceived as most familiar. This suggests that the experimental paradigm might be worth revisiting with a larger population.

In Chapter 4, participants had to make a Yes/No decision for each face in the lineup because we wanted to minimise extraneous cognitive load by minimising additional memory demands of remembering a number. However, in our pilot experiment, even though participants were explicitly told not to select Y for more than one face, several participants made multiple identifications within a single lineup presentation. This meant that we could not use their data, but it was useful to see how many participants had issues with making multiple false alarms. The results indicated that many of them identified the first face in the lineup. This was one of the main reasons that we included the practice session for the main experiment. The practice session helped reduce multiple false alarms considerably, reduced the number of participants who chose the first face in the lineup (as well as subsequent faces) to zero, and indicated that a practice session might also be a good idea in police procedures.

In Chapter 5, we followed UK police procedures by numbering the faces in the lineup, and asked participants to remember the number of a face if they wanted to identify it at the end of the lineup. However, one participant said afterwards that she picked the wrong number by mistake as she had forgotten the correct number, which made us wonder how many others had also forgotten the number but not told us. This made us consider the reliability of this type of procedure. However, it adds further weight to the use of pupillometry in police procedures, as pupillary responses can be independent of

decision responses (Heaver & Hutton, 2011), which is useful if an eyewitnesses chooses the wrong number by mistake.

In Chapter 6, we took the experiment to Brighton Pier, used a portable eye-tracker, and the target was a person the eyewitnesses had physically met rather than someone from a mock crime video. The main issues were related to conducting an experiment during a live event in terrible weather. Due to the weather, the Booster Ride that we were using closed early, our target had to wear weather-proof clothing and to seek shelter when possible, and very few people wanted to go on the ride in the first place. This meant that our participants were all people who were keen to go on the ride despite the weather, so we had very few participants who rated themselves as anxious. The lack of participants also meant that we did not have enough data to test the RKG satisfactorily. However, the experience gained from the issues with the previous experiments prepared us to cope with these adverse conditions. Our data that supported that of the previous experiments, and the experience of conducting an experiment that could not be piloted in situ was invaluable. We also have the experiment and equipment to hand so that we can re-test with more participants on another occasion.

7.1.4. Future Directions

Face learning - our research showed that the initial stages of face learning occurred gradually, but we concluded that the experiment was probably too easy for most people. It was also complicated by many variables. Therefore, it would be interesting to test a simplified (fewer variables) but more difficult version of this experiment with three groups of people: super-recognisers, typical face processors, and prosopagnosics, to see whether learning patterns and pupillary responses shed any light on the deficits of prosopagnosics. It would also be interesting to see what would occur if familiarisation

was continued for longer or if there were a gap between the familiarisation and learning stages, to clarify the effects of more exposure or gaps in learning on the formation of robust representations.

Understanding this might help to clarify some of the issues in prosopagnosia and introduce measures to address them. For instance, if it were found that reaching a threshold of exposure was key to learning a face, and that the poor face recognition skills made prosopagnosics either rely on alternative cues to recognition (e.g. voice, movement, context etc.) or withdraw from social interaction (Dalrymple et al. 2014), technology that increased exposure to socially-important faces to aid learning could be introduced. Similarly, if the length of gaps between viewing a face affected learning, technology could be introduced to bridge or create optimum gaps to improve learning. An app with timely notifications containing video clips of socially-important faces could address both exposure and viewing gaps if they were found to affect face learning.

Theoretical constructs (cognitive load, cognitive engagement, and memory strength) - our research in this area was largely inconclusive. The summary of our data was as follows: when learning novel faces in a short experiment, it appears that a gradual reduction in cognitive load was the most likely contributor to the pupillary changes, but we could not rule out that they might have occurred due to a decline in cognitive engagement. The results of Chapter 3 could have been explained by any of the theories, but we favoured a cognitive engagement account. It seems that pupillometry reflects fluctuations in mental effort, engagement, and memory, and our early experiments failed to test for one while controlling for the other, or to explore how they interact. Specifically, it appears that cognitive load and cognitive engagement might often co-occur and that pupillometry might not be the way to tease them apart.

It was only in the final half of the thesis, where we compared pupil size with explicit ratings of memory strength (RKG) that we concluded that memory strength offered a plausible account of the pupillary effects in a lineup. The addition of the RKG rating (that measured participants' belief in their memory) suggested that the pupillary responses in the lineup experiments reflected memory strength. This demonstrated that combining the pupillary responses with a self-rating measure was one way to clarify the pupillary results. Similar approaches could be worth investigating in terms of cognitive load (where participants' pupillary responses could be compared to self-rating of mental effort), and with cognitive engagement (where they could be compared to participants' self-rating of engagement). This could be particularly effective if an experiment could be designed that manipulated cognitive load while controlling for engagement or vice versa.

However, if we think of the lineup as being (metaphorically) a type of memory search task, it is apparent that the pupillary responses might also have been influenced to some extent by the different task demands between the chapters. In Chapter 4, the Y/N task probably encouraged a "self-terminating" strategy (which is when a person searching for an item stops processing items that are seen after an identification has been made), and this was reflected by the pupil sizes getting smaller once the target had been presented. However, Chapters 5 and 6 probably encouraged an "exhaustive" strategy (which is where people process each item and make an identification decision afterwards), as pupils remained a constant size throughout each lineup (apart from when identifiers viewed the target and their pupils dilated) (see Körner et al., 2014; Orzechowski, Nęcka, & Balas, 2016; & see e.g. Van Zandt & Townsend, 1993, for a review of self-terminating and exhaustive strategies). Thus, it is worth testing the pupillary responses in a lineup paradigm *without* asking participants also to make a conscious decision, to see whether the pupils respond to the target in the absence of any

conscious decision task. Nevertheless, the pupillary responses to the target in identifiers were dramatic whatever the task, adding weight to the idea that they reflected a response to the target that was *independent* of the conscious identification decisions that people made.

Forensic applications - our research indicated that pupillometry could provide an additional and practical tool to police procedures for measuring eyewitness credibility. Our final experiment on Brighton Pier demonstrated that pupillometry can also be effective in a portable device, but we were hampered by the weather. It would be worthwhile to use this experiment as a pilot for a series of other experiments, testing the effects of estimator variables on identification, including anxiety, alcohol, age, race, delay, exposure and so on. These experiments could also be extended to examine whether pupillometry could be viable with simultaneous lineups (although this seems unlikely), and testing locations (such as testing participants in their own homes, in the police station, or online). It would also be worth considering new technology to make it easier for people to use and more portable, such as an app.

For instance, an app that could calculate pupil size irrespective of head movements (by calibrating the distance between the pupils, to account for changes in distance between the eyes and the screen), or pupillometry goggles such as a version of the Tobii Pro (n.d.) (that could fix the distance and control for luminance levels) would allow for pupillometry lineups to be conducted in eyewitnesses' or victims' own homes. This could be done on a portable and familiar piece of equipment such as an iPad (n.d.). This would mean that the stress involved in viewing a lineup could be kept to a minimum. The advantage of using pupillometry, and the pupillary measure that we used (calculating the mean pupil size for each trial and converting each one into a percentage of the overall

pupil size range) means that other eye-tracking recordings (such as fixations, blinks and gaze patterns) would not be needed, so the software could be simple to create, and simple to score and analyse. The scope of this line of research is extensive and innovative, and further research could also provide enough data for meta-analysis, clarifying the findings of our previous research.

In short, while there were issues with the experiments in this thesis, the broad research set the scene for several lines of research measuring face processing with pupillometry. Pupillometry revealed nuanced effects that were independent of decision responses, so more research for expansion, clarification, and specialisation is warranted.

7.2. Conclusion

The research in this thesis provided novel insight into four main issues.

It used a novel approach to clarify differences between familiar and unfamiliar face processing. While more is now understood about the difference between familiar and unfamiliar faces in terms of processing, progress has been slow (Burton, 2013), and there remains much that requires further investigation. It seems that pupillometry shows some promise in this respect, as it can reflect subtle fluctuations in cognitive processing that behavioural measures (such as decision responses) fail to detect.

Face learning is one area that remains poorly understood, although some progress is now being made. The research in this thesis provided insights to how initial face learning occurs, weighting the previous evidence in favour of a gradual process, at least in the early stages of learning. This supports the idea that faces gradually become more robustly represented as different views of them are seen. This occurs quickly, well within

the confines of an experiment, although it is doubtful that experimentally-learnt faces can be represented as robustly as personally-familiar faces.

We evaluated the contributions of cognitive engagement, cognitive load, and memory strength to face processing. While we did not reach certainty, when cognitive engagement was controlled by presenting participants with novel faces that were learnt during an experiment, we concluded that cognitive load decreased as learning outcomes improved, and that unfamiliar faces imposed a higher load on mental resources than familiar ones, although the differences in pupil size were small (approximately 2%). However, when experiments contained faces that differed in terms of social importance to participants, any small differences in cognitive load appeared to be masked by larger ones associated with cognitive engagement. It may be that pupillometry is inadequate to tease apart cognitive load and cognitive engagement, or that they often co-occur, and we did not come to any strong conclusions about their contribution to face recognition in our experiments. However, pupillary changes were dramatically larger in the lineup experiments (32% in Chapter 4) than they had been in Chapters 2 and 3. We concluded that memory strength best accounted for these changes, as they only occurred in participants who identified the target in each experiment *as they were looking at them*, suggesting that they had remembered them. Thus, we were confident in the lineup chapters that memory strength best accounted for the pupillary responses.

As far as we are aware, our research was the first to use pupillometry in forensic settings, and our results suggests that it looks promising in this field. Not only did our pupillary results suggest that pupillary changes only occurred in participants who identified the target as they were looking at him, but we were able to use pupil sizes to predict identification responses. We were also able to conclude that the pupillary changes

only occurred in participants who recognised the target as they were looking at him, rather than being the consequence of making an identification. Finally, pupillary responses even appeared to indicate that some participants had implicitly recognised the target despite making no identification. Therefore, pupillometry has the potential both to reduce the rates of wrongful convictions and wrongful non-convictions. This research is forensically important, as pupillometry is the only current measure that can reliably reflect implicit recognition. It is also less likely to be contaminated by the processes required to make a conscious decision than the measures currently used in lineups. Therefore, pupillometry could provide an additional tool for police procedures.

The research in this thesis set the scene for several lines of research measuring face processing with pupillometry, as it revealed subtle effects that were unobtainable with decision responses, and showed that pupillometry could potentially have a promising role in real-world settings.

References

Ayres, P., & Paas, F. (2012). Cognitive Load Theory: New Directions and Challenges: Cognitive load theory: new directions. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brédart, S. (2003). Recognising the usual orientation of one's own face: The role of asymmetrically located details. *Perception*, *32*(7), 805–811. https://doi.org/10.1068/p3354

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364.

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology*, *53*(12), 1823–1835. https://doi.org/10.1111/psyp.12770

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces: Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*(1), 41-56.

Dalrymple, K. A., Fletcher, K., Corrow, S., das Nair, R., Barton, J. J. S., Yonas, A., & Duchaine, B. (2014). "A room full of strangers every day": The psychosocial impact of developmental prosopagnosia on children and their families. *Journal of Psychosomatic Research*, *77*(2), 144–150. https://doi.org/10.1016/j.jpsychores.2014.06.001

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological review, 115*(2), 426-446.

Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: five days to execution and other dispatches from the wrongly convicted* (1st ed). New York: Doubleday.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science, 21*(2), 90-95. https://doi.org/10.1177/0963721412436811

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

The Innocence Project (n.d.), retrieved 18th May, 2018, from https://www.innocenceproject.org/causes/eyewitness-misidentification/

iPad (n.d.), retrieved May 25th 2018, from https://www.apple.com/uk/ipad/

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Kircher, T. T. J., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P. J., Bullmore, E. T., … David, A. S. (2001). Recognizing one's own face. *Cognition*, *78*(1), B1–B15. https://doi.org/10.1016/S0010-0277(00)00104-9

Körner, C., Braunstein, V., Stangl, M., Schlögl, A., Neuper, C., & Ischebeck, A. (2014). Sequential effects in continued visual search: Using fixation-related potentials to compare distractor processing before and after target detection. *Psychophysiology*, *51*(4), 385-395.

Laeng, B., & Falkenberg, L. (2007). Women's pupillary responses to sexually significant others during the hormonal cycle. *Hormones and Behavior*, *52*(4), 520–530. https://doi.org/10.1016/j.yhbeh.2007.07.013

Miller, M. K., & Bornstein, B. H. (Eds.). (2013). *Stress, Trauma, and Wellbeing in the Legal System*. Oxford, UK: Oxford University Press. DOI: 10.1093/acprof:oso/9780199829996.001.0001

Moreno, R., & Park, B. (2010). Cognitive Load Theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 9–28). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.003

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic bulletin & review, 23*(5), 1316-1340. https://doi.org/10.3758/s13423-015-0982-5

Orzechowski, J., Nęcka, E., & Balas, R. (2016). Task conditions and short-term memory search: two-phase model of STM search. *Polish Psychological Bulletin*, *47*(1), 12-20.

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory: Pupil size and recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Prehn, K., Heekeren, H. R., & van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243. https://doi.org/10.1016/j.ijpsycho.2010.10.014

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*(8), 1217–1223. https://doi.org/10.1111/psyp.12668

SR Research (n.d), retrieved 19th May, 2018, from https://www.sr-research.com/products/eyelink-portable-duo/

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262–274. https://doi.org/10.1007/s10979-010-9236-2

Sweller, J. (2010). Cognitive Load Theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 29–47). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Tacikowski, P., & Nowicka, A. (2010). Allocation of attention to self-name and self-face: An ERP study. *Biological Psychology*, *84*(2), 318–324. https://doi.org/10.1016/j.biopsycho.2010.03.009

Tobii Pro (n.d.), retrieved May 25th 2018, from https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

Van Zandt, T., & Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics*, *53*(5), 563-580.

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360-376.

Wells, G. L., & Olson, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wixted, J. T., Read, D.J., & Lindsay, S.D. (2016). The Effect of Retention Interval on the Eyewitness Identification Confidence–Accuracy Relationship. *Journal of Applied Research in Memory and Cognition*, *5*(2), 192–203. https://doi.org/10.1016/j.jarmac.2016.04.006

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Appendices:

Appendix 1. The RKG statements.

How confident are you in your response?

If you made an identification, go to A. If you made no identification, go to B.

A.

If you feel confident that you identified perpetrator because you could remember his face, select R (remember).

If you had a sense of knowing (gut feeling) that you identified the perpetrator without actually remembering the perpetrator's face, select K (Know).

If you were unsure, but felt compelled to identify someone, select G (Guess). For example, if you wanted to select a face rather than not select a face, or because the face was more similar to the perpetrator than the other faces

R K G

B.

If you feel confident that the perpetrator was not present because you remember his face, select R (remember).

If you had a sense of knowing (gut feeling) that the perpetrator was not present, without actually remembering the perpetrator's face, select K (Know).

If you were unsure, but all the video clips had been shown without you making a decision, select G (Guess).

R K G